From LABORATORY MEDICINE
Karolinska Institutet, Stockholm, Sweden

ELUCIDATING REGULATORY ELEMENTS: STUDIES IN CHRONIC LYMPHOCYTIC LEUKEMIA AND MULTIPLE MYELOMA

Ayla De Paepe

Karolinska Institutet

Stockholm 2019

Published by Karolinska Institutet.
Printed by Eprint AB 2018

ISBN 978-91-7831-311-2

Elucidating regulatory elements: studies in Chronic Lymphocytic Leukemia and Multiple Myeloma

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Ayla De Paepe

Principal Supervisor:
Robert Månsson, Assistant professor
Karolinska Institutet
Department of Laboratory Medicine

Co-supervisor(s):
Pelin Sahlén, Assistant professor
KTH Royal Institute of Technology
Science for Life Laboratories

Julian Walfridsson, Assistant professor
Karolinska Institutet
Department of Medicine

Lennart Hammarström, Professor
Karolinska Institutet
Department of Laboratory Medicine

Opponent:
Christopher Benner, Assistant professor
University of California San Diego
Department of Medicine

Examination Board:
Richard Rosenquist Brandel, Professor, MD
Karolinska Institutet
Department of Molecular Medicine and Surgery

Ann-Kristin Östlund Farrats, Professor
Stockholm University
Department of Molecular Biosciences

Markus Ringnér, Bioinformatician, Docent
Lund University
Department of Biology

Genome: Bought the book; hard to read.

– Eric Lander

To my family and all the joys of life

ABSTRACT

With next generation sequencing taking center stage in genetic and epigenetic research, its applications and challenges are many. This work revolves around the application of bioinformatics in different contexts: basic research in the understanding of diseases (biology), the effect of treatment on the target cells (clinics) and the assessment of a new wet-lab method (lab).

Biology Two studies fall under this topic, one on chronic lymphocytic leukemia, the other on multiple myeloma. Many coding mutations and chromosomal aberrations have long been identified in both diseases, yet they are only present in subsets of patients, and so it is puzzling that all this diversity results in a single diagnosis. We hypothesized that instead of a common genetic background, they might present with a common epigenetic background. For this we aimed to collect paired RNA-seq, histone ChIP-seq and ATAC-seq for patients and healthy controls, and had the following specific hypotheses:

1. Using H3K4me2 and H3K27Ac, we will be able to identify the regulatory elements altered between health and disease
2. By looking at the interplay between those regulatory elemets, RNA-seq, ATAC-seq and database information, we will be able to describe the aberrant regulation in terms of transciption factors, regulatory elements and their target genes.

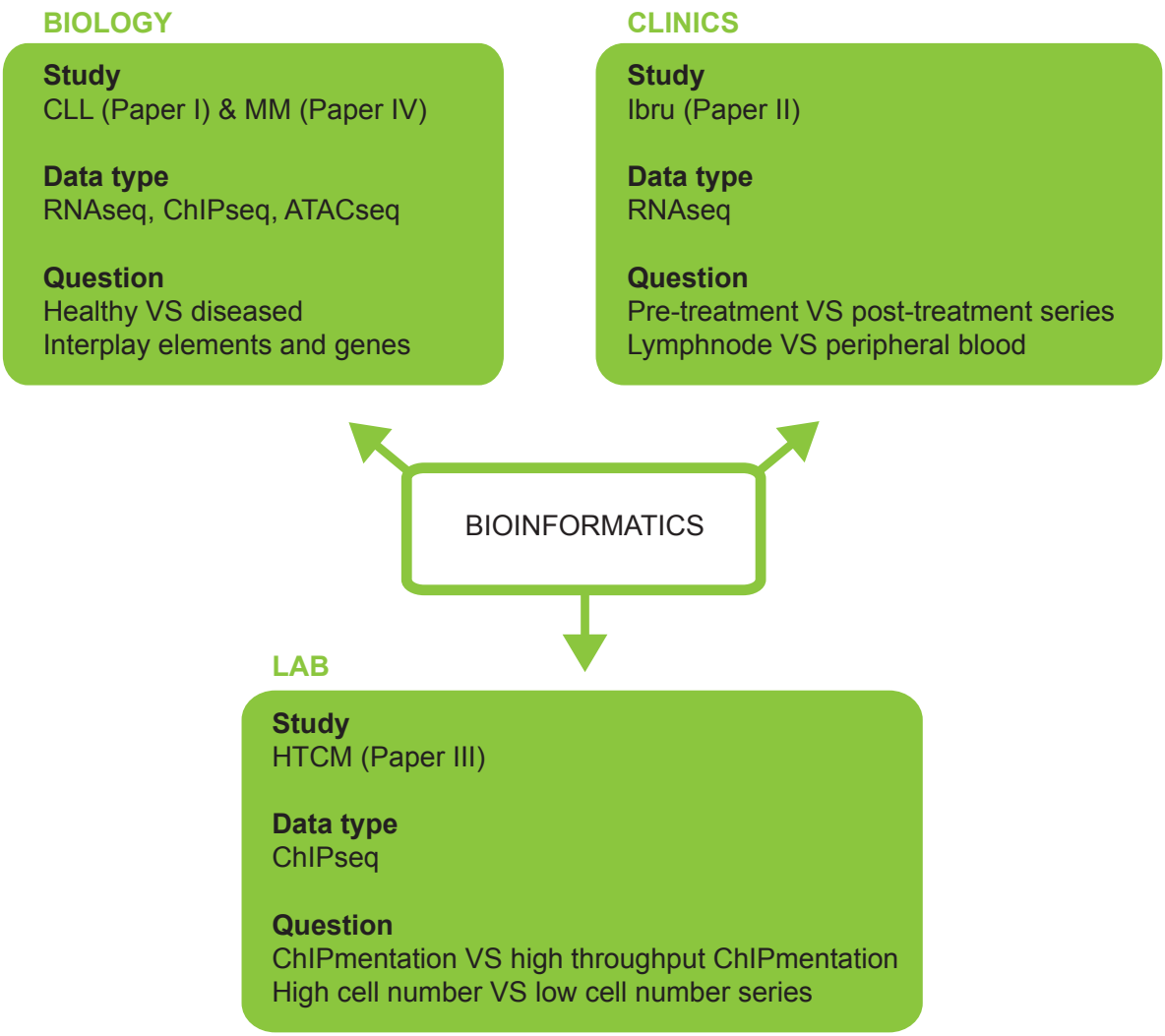
Clinics Ibrutinib is a novel drug used in chronic lymphocytic leukemia treatment. Though it is shown to be beneficial to many patients, a lot of the early effects the treatment has on the malignant cells, especially in different parts of the body, are still unknown. We hypothesized that relevant changes in blood and lymph nodes would be visible soon after treatment, and collected RNA-seq data from both compartments, in additions to plasma values of inflammatory cytokines. We had to following specific hypotheses:

1. Early changes are visible, maybe even just hours after first treatment.
2. There will be differences in treatment effect between blood and lymph node compartments.

Lab ChIP-seq is a very useful method to look at proteins bound to DNA, which, depending on the protein checked, can give a multitude of information. Yet, the amount of cells needed to perform these experiments is high, even in improved protocols like ChIPmentation. We hypothesized that the ChIPmentation protocol could be optimized, and that reducing time and steps would yield better data. For this we performed ChIP-seq with the original ChIPmentation protocol and with our adaptation, testing the following specific hypotheses:

1. Our version, high-throughput ChIPmentation, will perform equally well as the original method when high cell numbers are used, but be faster.
2. When low cell numbers are used, our method will give better results, as it involves less loss of material.

My contribution lies in the development and execution of bioinformatics, pipelines, data handling etc, to test these hypotheses, which also includes discussions and planning of projects, samples and feasibility.



LIST OF SCIENTIFIC PAPERS

Four papers are included in this work. For ease of interpretation they will be referred to using their abbreviations.

* Contributed equally
** Contributed equally, corresponding

I. **CLL**
Chronic lymphocytic leukemia is underpinned by a common gene regulatory network despite diverse genetic abnormalities

Ayla De Paepe*, Aleksandra Krstic*, Charlotte Gustafsson, Ruth Clifford, Marcin Kierczak, Kristina Sonnevi, Xaoze Li, Shabnam Kharazi, Minna Suomela, Kenian Chen, Malin Larsson, Dag Ahrén, Eva Hellqvist, Minna Taipale, Yin Lin, Eva Kimby, Anna Schuh, Robert Mansson

In manuscript

II. **Ibru**
Ibrutinib induces rapid down-regulation of inflammatory markers and altered transcription of chronic lymphocytic leukaemia-related genes in blood and lymph nodes

Marzia Palma*, Aleksandra Krstic*, Lucia Pena Perez, Anna Berglöf, Stephan Meinke, Qing Wang, K. Emelie M. Blomberg, Masood Kamali-Moghaddam, Qiujin Shen, Georg Jaremko, Jeanette Lundin, **Ayla De Paepe**, Petter Höglund, Eva Kimby, Anders Österborg*, Robert Månsson* and C. I. Edvard Smith*

British Journal of Haematology, 2018, 183, 212-224, doi: 10.1111/bjh.15516

III. **HTCM**
High-Throughput ChIPmentation: freely scalable, single day ChIPseq data generation from very low cell-numbers

Charlotte Gustafsson, **Ayla De Paepe**, Christian Schmidl and Robert Månsson

Accepted for publication in BMC genomics

IV. **MM**
Active enhancer and chromatin accessibility landscapes chart the regulatory network of primary multiple myeloma

Yi Jin*, Kenian Chen*, **Ayla De Paepe***, Eva Hellqvist, Aleksandra D. Krstic, Lauren Metang, Charlotte Gustafsson, Richard E. Davis, Yair M. Levy, Rakesh Surapaneni, Ann Wallblom, Hareth Nahi, Robert Mansson** and Yin C. Lin**

Blood, 2018, 131, 2138-2150, doi: 10.1182/blood-2017-09-808063

CONTENTS

LIST OF ABBREVIATIONS

Ab	antibody
aCNV	acquired genomic copy number variation
BCR	B-cell receptor
bp	base pairs
CLL	chronic lymphocytic leukemia
CM	ChIPmentation
d2	two days / day two
d29	29 days / day 29
FC	fold change
FDR	false discovery rate
FP	footprint
GC	germinal center
GTF	general transcription factor
ac	H3K27ac
me2	H3K4me2
HDAC	histone deacetylase
HDACi	histone deacetylase inhibitor
HTCM	high-throughput ChIPmentation
ibru	Ibrutinib
Ig	immunoglobulin
LN	lymph node
MB	memory B-cell (in multiple myeloma paper context)
MemB	memory B-cell
MM	multiple myeloma
NaïveB	naïve B-cell
nsMemB	non-switched memory B-cell
PB	peripheral blood, or plasma blast (in multiple myeloma paper context)
PC	plasma cell (in multiple myeloma paper context)
PolII	RNA polymerase II
PCR	polymerase chain reaction
SE	super-enhancer
sMemB	switched memory B-cell
SNP	single nucleotide polymorphism
TAD	topologically associated domain
TF	transcription factor

TFBS	transcription factor binding site
TSS	transcription start site
VAF	variant allele frequency

1 B-CELLS

All studies included in this work are related to B-cells, be they healthy or diseased, primary or cell lines. Hence, it seems only reasonable to give them a little introduction.

1.1 ROLE

B-cells are lymphocytes (**Figure 1**) responsible for antibody (Ab) production. Abs bind to antigens, any recognizable structure, and in this way flag unknown agents to the rest of the immune system. In order to react with as many foreign antigens as possible, the B-cell population in a healthy individual is made up of thousands of clones. All of these clones are genetically different and produce slightly different Abs. The genetic differences observed in B-cell clones are accomplished by recombination of the immunoglobulin (Ig) regions, the genetic regions that code for Ab parts. Thus, in contrast to most other cell types, which all have the exact same genome, B-cells undergo genome editing.

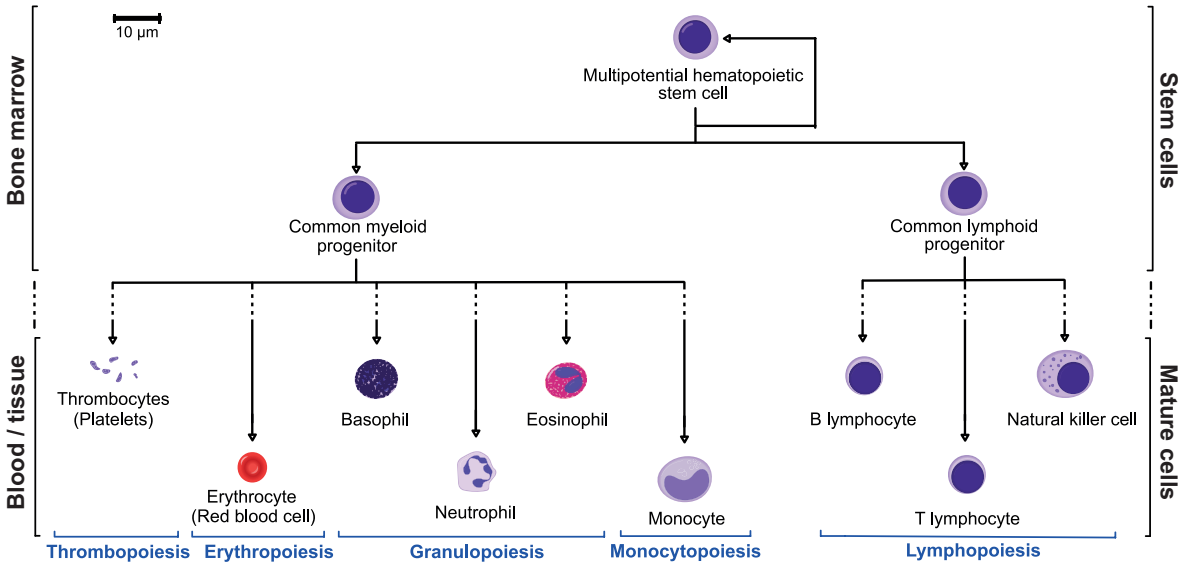


Figure 1: Simplified overview of the hematopoietic system in humans.
Figure adapted from creative commons illustration by A. Rad[1].

1.2 B-CELL ACTIVATION

When a mature, naïve B-cell (NaïveB), binds an antigen by use of its B-cell receptor (BCR), the cell becomes activated (**Figure 2**).

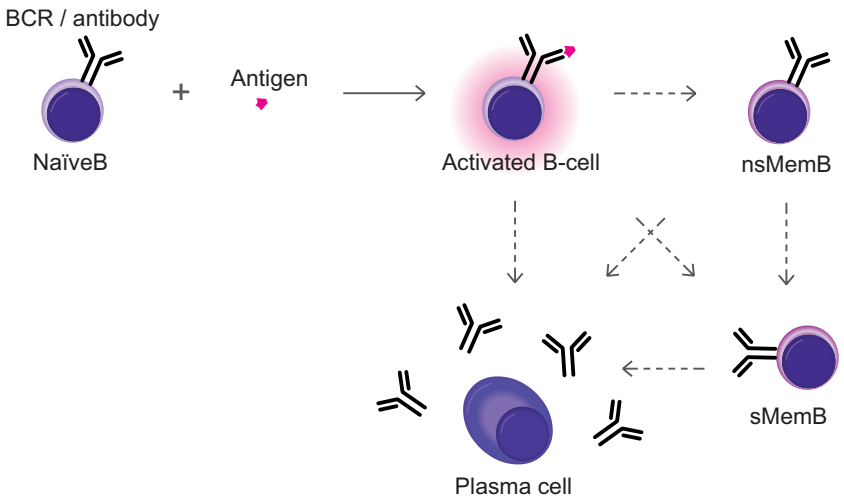


Figure 2: B-cell activation. NaïveB = mature, naïve B-cell, BCR = B-cell receptor, nsMemB = non-switched memory B-cells, sMemB = switched memory B-cells. Figure used elements from creative commons illustration by A. Rad[1]

Activated B-cells will make their way to the germinal center (GC) in the lymph nodes. Here they undergo GC reaction, resulting in somatic hyper mutation of the Ab part regions in their genome. The goal is to create a B-cell clone that reacts even better with the antigen that caused the activation, and thus perfect the immune reaction. A part of these perfected B-cells will become plasma cells, which are bigger, and optimized for producing high numbers of secreted Abs. These circulating Abs will bind to the agent that triggered the immune reaction, cluster together and flag them for destruction by other immune cells. Other activated B-cells will become memory B-cells, which will stay in the body and, in case of a second infection with the same agent, will be ready for a fast, targeted immune reaction. These memory B-cells can still have IgM and IgD Abs, as the original mature B-cell had, or switch to IgE, IgG or IgA Abs, referred to as non-switched memory B-cells (nsMemB) and switched memory B-cells (sMemB) respectively[2].

2 TRANSCRIPTIONAL REGULATION

In order for cells to function properly, they need strict regulation of what genes are expressed, and when and where this happens.

2.1 REGULATORY ELEMENTS

Eukaryotic, protein-coding genes generally have two types of regulatory DNA elements: promoters and distal regulatory elements. The later include elements such as enhancers and insulators[3]. In order for these to fulfill their roles at the right time and place, they interact with transcription factors (TFs), which will be discussed in a later section.

2.1.1 Promoters

Promoters are regulatory elements located right at the transcription start site (TSS) of the gene they regulate. Promoters are usually made up of a core promoter and a proximal promoter[3]. The core promoter is the region that stretches about 35 base pairs (bp) up- and/or downstream of the gene's TSS and is the docking site for basic transcription machinery like RNA polymerase II (PolII)[4]. The region directly upstream of the core promoter is the proximal promoter, which is made up of the regulatory elements that fall up to a few hundred bp away from the core promoter[3]. These regulatory elements contain binding sites for TFs regulating the expression of the gene in question (**Figure 3A and B**).

2.1.2 Distal regulatory elements

Enhancers are regulatory regions containing a set of TF binding site collections that can increase the transcription of a gene, which makes them very similar to proximal promoters[3]. The biggest difference is the distance at which they operate. While promoters are just upstream of the TSS, enhancers can be hundreds of kilobases away and located anywhere in the non-coding part of the genome. This means they can be up- or downstream of the gene, or even within a gene body. However, to exert its regulatory effect, the enhancer needs to be brought into proximity of the promoter of its target gene, which is done by looping out the DNA that separates them (**Figure 3C**). The result is that both are still far away in linear sequence, but close in 3D space. Though the direction of the whole enhancer is not critical for its function, the enhancer might only be able to regulate its target gene if the different parts that make up the enhancer are correctly positioned relative to one another.

2.1.3 Super-enhancers

In addition to these regular enhancers, usually a few kilobases in size[5], larger regions with multiple regular enhancers clustering together in close proximity have been identified and termed super-enhancers (SE). Though there is still discussion on whether they are a separate type of regulatory elements or rather a collection of regular enhancers[6], SE are often linked to cell type specific genes, making them important for cell identity[7,8]. SE have also been associated to disease[8,9], and the fact that they are easily disrupted might make them interesting drug targets[10,11].

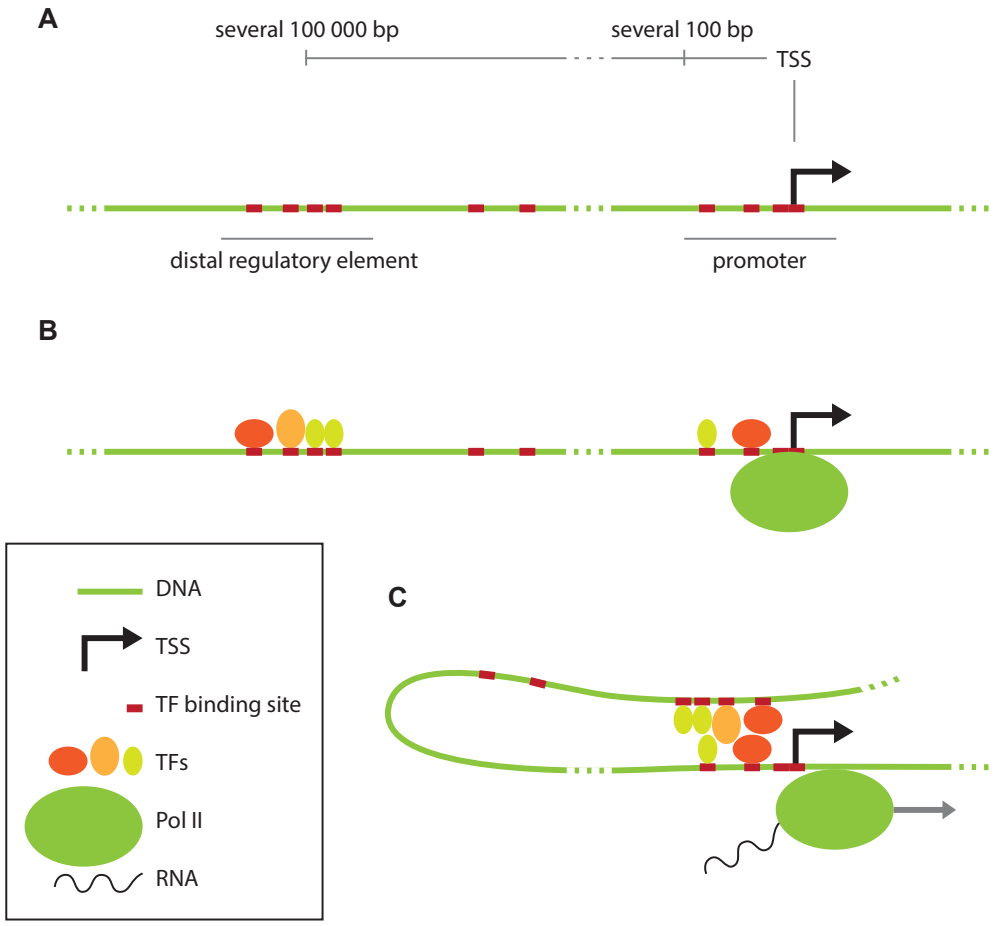


Figure 3: Representation of DNA, regulatory elements and transcription factors.

2.2 TRANSCRIPTION FACTORS

Transcription factors (TFs) are proteins that change the expression of genes. They can either enhance or repress expression and as a result are respectively called activators or repressors[5]. They can also be organized based on their way of interacting with regulatory elements[3]. General transcription factors (GTFs) interact directly with the core promoter of a gene, and are the minimum TFs needed for basal transcription. They include RNA polymerase II and other factors that make up the pre-initiation complex. The resulting basal transcription is not particularly high or regulated, but can be altered by activators and repressors. This second group of TFs directly binds DNA sequences of about 6-12bp, called motifs or binding sites, in enhancers and proximal promoters[3,5] (**Figure 3B**). They are classified into TF families based on the motif they recognize and the domain structure that binds it[12]. In addition to a DNA binding domain, activators also have an activation domain required for enhancing the transcription of their target gene[5]. Repressors on the other hand usually work by blocking activators or GTFs from binding their motifs. The last group of TFs is that of the co-activators and co-repressors. These TFs do not have DNA binding domains themselves, but interact with activators and repressors to regulate transcription[3].

The resulting set-up can be rather complex, and the correct regulation of a gene generally requires multiple TFs to exert their function in the right time and place. Though multiple models exist[5], this usually means the TFs need to be expressed and available at the right level, their motifs need to be free for binding and all interacting partners need to be present at that moment, in that location.

2.3 EPIGENETICS

The regulatory elements described above are not based solely on DNA sequence (**Figure 4**). After all, all cells in the body, with some exceptions like B-cells, have the exact same genome, yet perform very different functions. For this, genomic regions can be activated and deactivated, guiding regulatory elements and TFs depending on the cell's need. This is where epigenetics comes into play. In B-cells, for example, processes all the way from general hematopoiesis[13], to B-cell specific events like Ig recombination[14,15], GC reaction and transition to memory B-cells[16–18], have been linked to epigenetic changes or regulation.

2.3.1 Definition

Though the term "epigenetics" is frequently used, definitions of what it actually means vary widely. The oldest definition is made by Waddington, who based it on "epigenesis", being "the concept that an organism develops by the new appearance of structures and functions, as opposed to the hypothesis that an organism develops by the unfolding and growth of entities already present in the egg at the beginning of development (preformation)" [19]. As a result, in this definition epigenetics includes all molecular mechanisms "by which the genes of the genotype bring about phenotypic effects", or basically anything that influences gene expression.

A stricter definition by Riggs makes the limitation that these changes need to be heritable, making epigenetics "the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence" [20]. Originally this only included DNA methylation, but some authors have also included more concepts like prions [21], being heritably misfolded proteins.

More recently, intermediate definitions have been suggested, including the one by Bird, who considers epigenetics to be "the structural adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states" [22]. This definition includes DNA methylation, but also heritable or transient histone modifications, which might only come about in specific stages of a cell's existence.

This work uses the last definition, which is, notwithstanding the lack of consensus, at least a common use of the term epigenetics[23–25]. As DNA methylation is the bread and butter of epigenetics, and histone modifications are the main focus of the included papers, those will be described a bit further.

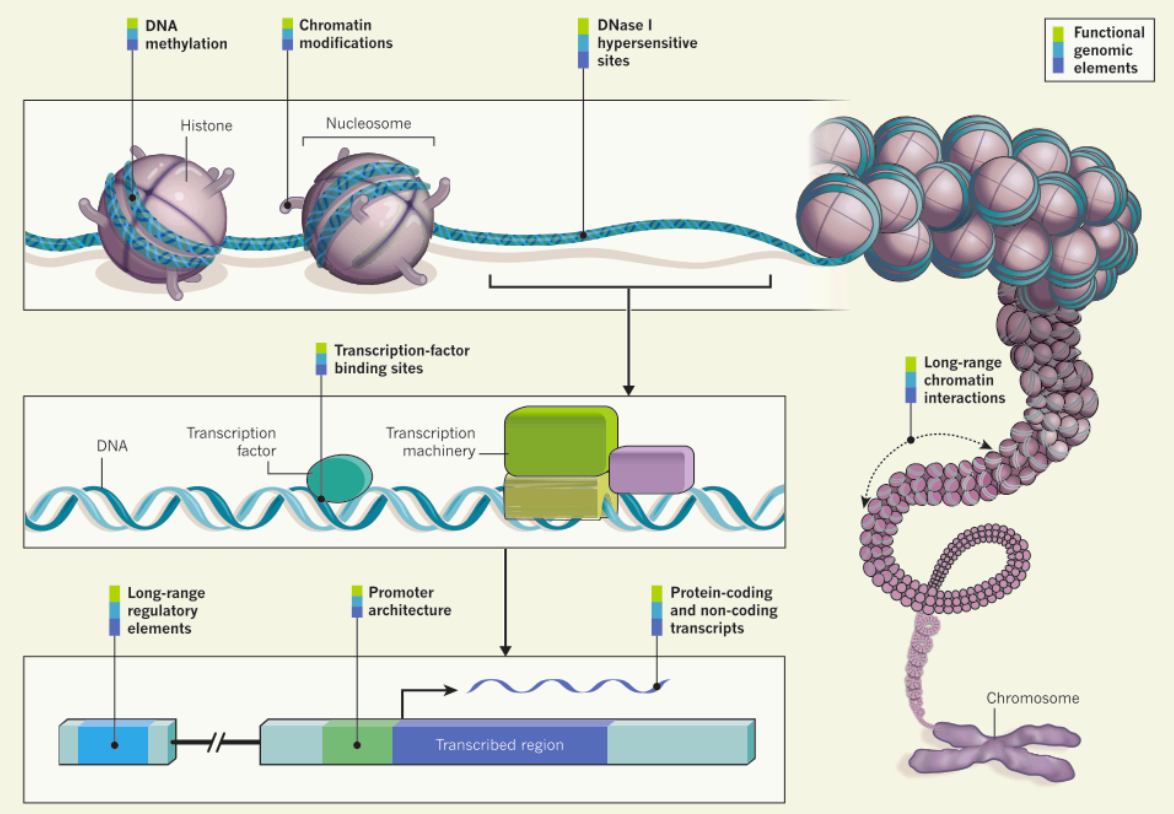


Figure 4: Overview of DNA organization, epigenetic marks and regulatory elements. Originally published by Ecker et al.[26], reproduced with permission from Nature.

2.3.2 DNA methylation

Cytosin (C) bases in the genome can be methylated by DNA methyltransferases. As a result, about 1% of the bases in the genome are 5-methylcytosine (5mC)[27]. This 5mC can spontaneously deaminate to a thymine (T) base[28]. It is speculated that this process is the basis for the observation that the genome generally has a higher AT than GC content. However, some regions in the genome are unmethylated while particularly GC rich. Due to their specific enrichment for a dinucleotide made up of C followed by G (CpG), these regions are called CpG islands. This specific DNA composition could destabilize the packing of the DNA in that region, making it more accessible for transcription[29]. CpG islands comprise about 1% of the mammalian genome[30] and harbor roughly 70% of TSS in humans[31,32].

Highly expressed genes tend to have unmethylated core-promoters and a methylated gene body[33], repressing incorrect transcription starting in the middle of a gene. When DNA at promoter or regulatory regions gets methylated, this interferes with proper expression and the associated gene is silenced[34]. However, the lack of DNA methylation does not necessarily mean a gene will be expressed. Both housekeeping genes, which are generally expressed, and tissue-specific genes can have their TSSs in unmethylated CpG islands, which would favor their expression, and still not be expressed in that moment and tissue[35].

2.3.3 Histone modification

In the eukaryotic genome, DNA is organized and compacted by winding around histones. A histone is an octamer, made out of four basic histone proteins: H2A, H2B, H3 and H4. Histone modifications are post-translational modifications to the tails of these histone proteins. There are at least twelve types of histone modifications, including methylation and acetylation, which can occur at a wide variety of residues within the histone tails[24]. As a result, the nomenclature of histone modifications is based on the type of modification and its location. For example, H3K27Ac refers to acetylation (Ac) on the lysine (K) residue that is in the 27th position in the tail of the histone protein H3. In addition to modifications to the four basic histone proteins, variants of these proteins have been observed, resulting in even higher diversity. The sum total of all these possible states has been termed the histone code[36].

The histone code exhibits general patterns, and distinct features in the genome display different set-ups in their histone code. Acetylation of histone tail residues is related to active regions of the genome and thus can be found on active enhancers and promoters. Methylation can be observed in either active or repressed regions, depending on the level of methylation and the exact residue involved[25]. For example, active promoters are shown to be enriched in H3K4me3 to a degree correlated to their expression[37] but are repressed when marked with H3K27me3 instead[38]. Looking at combinations of histone marks can give additional information. It has been shown that combining H3K4me2 and H3K27Ac marks can reduce false positives when predicting TF binding[39].

3 CHRONIC LYMPHOCYTIC LEUKEMIA & MULTIPLE MYELOMA

All blood lineages are challenged by cancers (**Figure 5**). Chronic lymphocytic leukemia (CLL) and multiple myeloma (MM) are part of the B-cell malignancies.

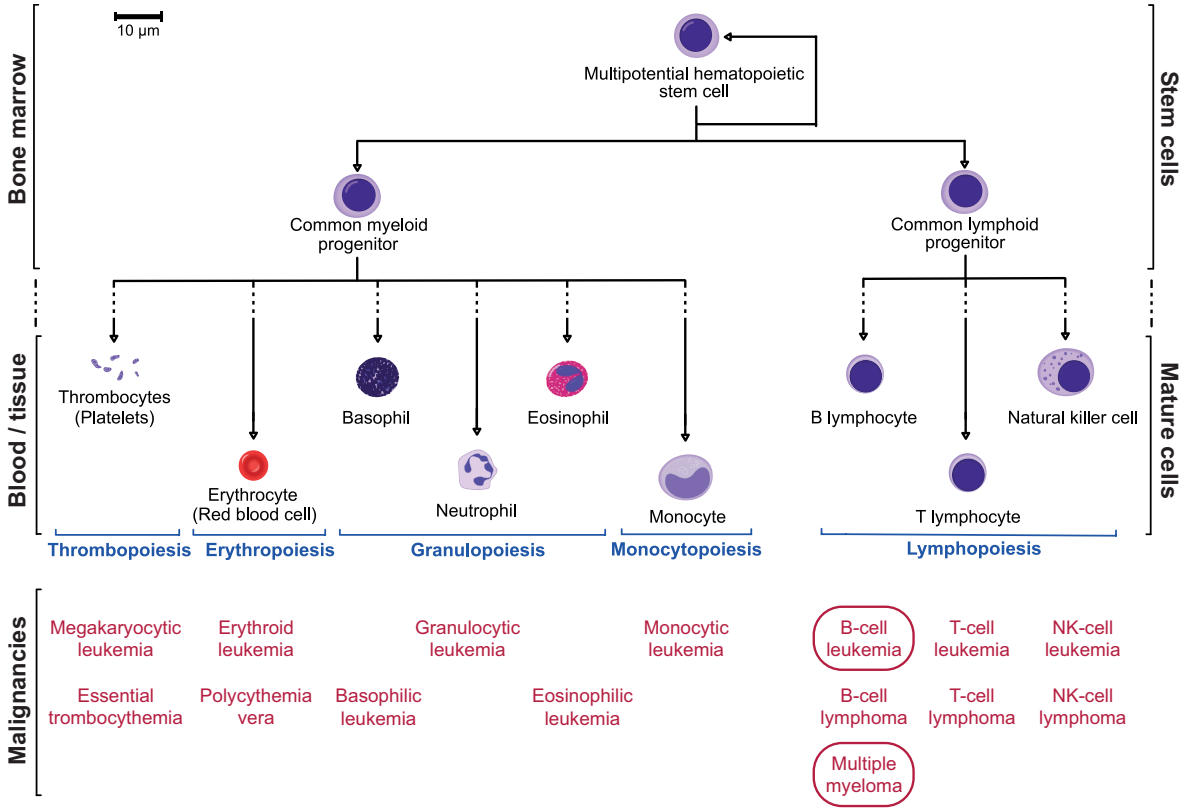


Figure 5: Simplified overview of the hematopoietic system in humans, with examples of malignancies in each lineage. Figure adapted from creative commons illustration by A. Rad[1].

Table 1: Estimated US cases in 2018 (by sex) for all cancers, lymphoma, multiple myeloma and leukemia[40]

	Estimated new cases			Estimated deaths		
	M & F	M	F	M & F	M	F
All sites	1,735,350	856,370	878,980	609,640	323,630	286,010
Lymphoma	83,180	46,570	36,610	20,960	12,130	8,830
Hodgkin lymphoma	8,500	4,840	3,660	1,050	620	430
Non-Hodgkin lymphoma	74,680	41,730	32,950	19,910	11,510	8,400
Myeloma	30,770	16,400	14,370	12,770	6,830	5,940
Leukemia	60,300	35,030	25,270	24,370	14,270	10,100
Acute lymphocytic leukemia	5,960	3,290	2,670	1,470	830	640
Chronic lymphocytic leukemia	20,940	12,990	7,950	4,510	2,790	1,720
Acute myeloid leukemia	19,520	10,380	9,140	10,670	6,180	4,490
Chronic myeloid leukemia	8,430	4,980	3,450	1,090	620	470
Other leukemia	5,450	3,390	2,060	6,630	3,850	2,780

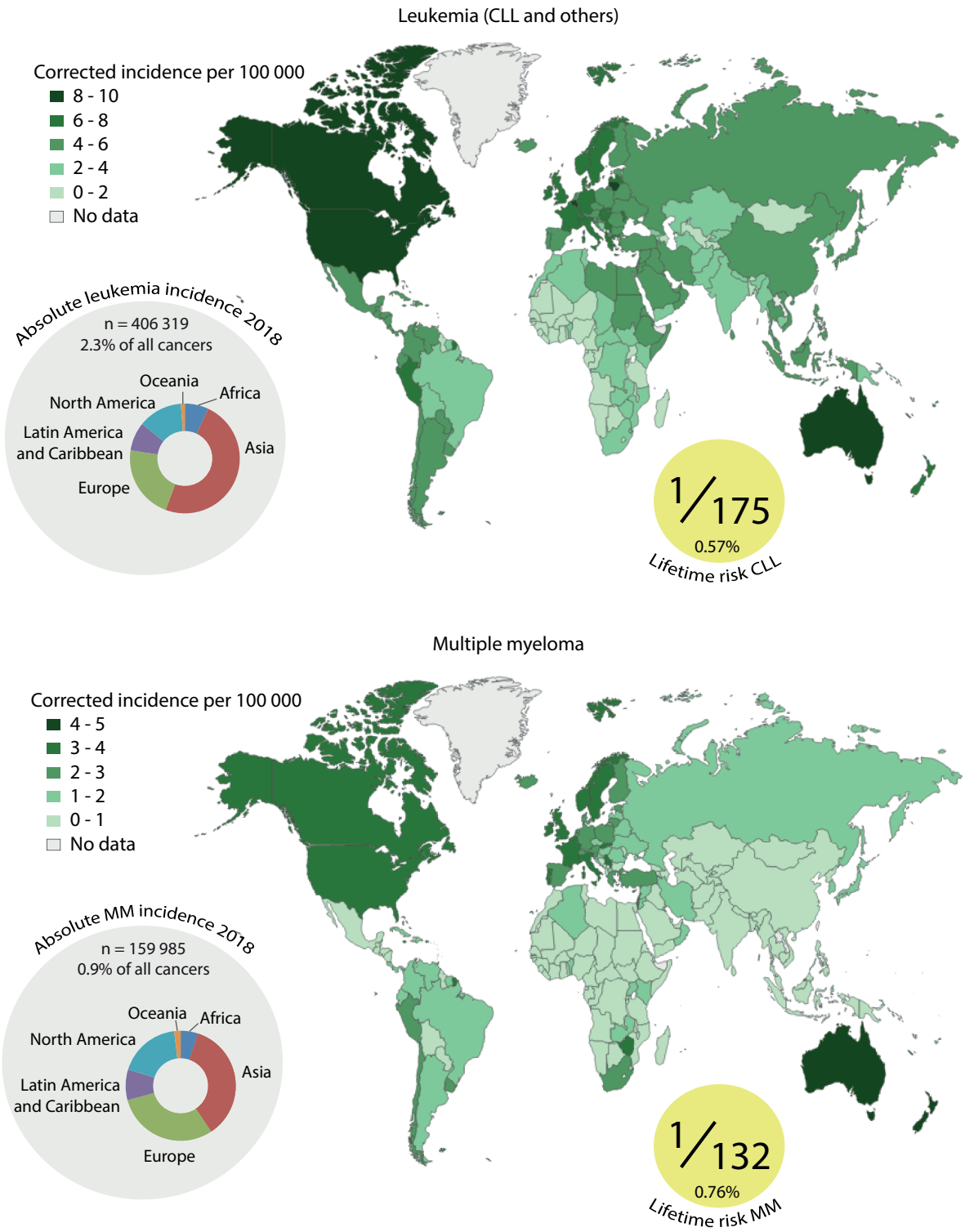


Figure 6: Estimated new cases of MM and leukemia (including CLL) in 2018 (incidence)[41,42]. World maps show estimated age-standardized incidence rates worldwide in 2018, including both sexes and all ages, and their visualization was made with Paintmaps[43]. Note that the scale for both world maps is different, but that the Leukemia one includes more than just CLL. Absolute incidence is from the same source, but not age-standardized. Lifetime risk estimates are US based, gathered from respective pages on the American Cancer Society website: CLL[44], MM[45].

3.1 CLL

CLL is the most common leukemia in adults in the western world, representing roughly 1/3rd of leukemia incidence and 1.2% of total cancer incidences in the United States[40] (**Table 1**). CLL incidence is higher in men than women, and multiple other risk factors are recognized (**Table 2**). Usually, CLL is diagnosed in a routine blood test, since it manifests as B-lymphocytes coming from one single B-cell clone accumulating in the peripheral blood, lymph nodes and bone marrow, even if no symptoms are apparent. However, small expansions of multiple B-cell clones are normal in elderly people, and do not need to result in CLL. If these small clones do continue to expand and select, this can turn into monoclonal B-cell lymphocytosis (MBL), and later on into CLL. The biggest distinction made within CLL is based on the level of mutations in the IGHV region of the Ab part genes. Patients with low mutational load in this region are called un-mutated CLL (U-CLL), while patients with high mutational load are called mutated CLL (M-CLL). This distinction is believed to be related to the B-cell of origin, and is not to be confused with the absence or presence of recurrent mutations in other regions of the genome.

Though CLL patients can go without symptoms for years, over time the expanding B-cell clone outcompetes the other blood cell lineages, resulting in anemia and infections that lead to the patient's death. Treatment is usually antibody therapy in combination with chemotherapy, but unless bone marrow transplantation is possible, CLL remains incurable. Commonly, a *watch and wait* approach is taken, delaying treatment as long as possible to avoid side effects and selective pressure of therapy risking future expansion of more aggressive clones.

3.2 IBRUTINIB IN CLL

As Ibrutinib is central in one of the studies included in this work, this section provides some more information on this drug. Though also used in other hematopoietic malignancies, its use in CLL was approved in 2014 in the US and Europe[46]. Initially this was only for previously treated patients, but by now it has been approved as monotherapy in first-line treatment[47]. Both as combination (HELIOS trial)[48] and monotherapy (RESONATE trials)[49,50], Ibrutinib has been shown to be more effective than standard treatment. Side effects are assessed to be tolerable, though alternatives might be preferred if one has specific other conditions[47].

Ibrutinib irreversibly binds, and thus inhibits, the Bruton's tyrosine kinase (BTK), which is one of the enzymes active in B-cell receptor signaling. Though healthy B-cells also use this pathway, Ibrutinib preferably kills CLL cells rather than healthy cells[51], and reduces tumor proliferation and load in blood, lymph nodes and bone marrow[52]. Because of this disruption of signaling, Ibrutinib induces sizable redistribution lymphocytosis, were cells resident in lymph nodes get released into the blood, resulting in a rise in absolute lymph counts in blood[53]. However, this seems to be harmless and reduces with further treatment.

3.3 MM

Like CLL, MM is most common in highly developed countries, representing about 0.9% of cancer incidences in the world[41,42] (**Figure 6**), but 1.8% of cancer incidences in the United States[40] (**Table 1**). Risk factors are shared with CLL (**Table 2**), again with higher incidence in men than women. While CLL probably originates in mature B-cells, somewhere around naïve and memory B-cells, MM originates somewhere around memory and plasma cells. Small clonal plasma cell expansions can grow into monoclonal gammopathy of undetermined significance (MGUS) and continue to MM.

Diagnosis can be based on end-organ damage, were patients might have pain in bones due to lesions, feel exhausted due to anemia or present with kidney failure[54]. Even without damage, the presence of expanded clones, accumulation of monoclonal protein levels and biomarkers can be sufficient for diagnosis.

Table 2: Overview of some of the risk factors and symptoms observed in CLL and MM.

	CLL	MM
Risk factors	• Advanced age (>65) • Male • White-American (vs other American backgrounds)[55] • Family with CLL[56] • Being a farmer (insecticides)[57] • Exposure to Agent Orange[58]	• Advanced age (>65) • Male • African-American (vs other American backgrounds)[59] • Family with MM[60] • Being a farmer (unknown exposure)[61]
Symptoms	Often none, otherwise e.g.: [44,62] • Anemia => tired,… • Leukopenia => infections,… • Enlarged spleen/liver => feeling full • Swollen lymph nodes	CRAB features: [54,63,64] • hyperCalcemia => dehydration,… • Renal failure => weakness,… • Anemia => tired,… • lytic Bone lesions => pain,…

3.4 GENETIC CHANGES IN CLL & MM

Both CLL and MM present a range of genetic aberrations. They differ not only compared to healthy individuals, but also between each other, and between individuals with the same disease. CLL displays a number of recurrent mutations and acquired genomic copy number variations (aCNVs), all present in less than half of patients, and the genetic diversity in MM

is a lot higher still. Charting changes is not only interesting for understanding the disease, but bears great potential for individualized medicine.

3.4.1 SNPs with increased risk

Single nucleotide polymorphisms (SNPs) are normally occurring differences in genetic code between individuals. These variations are the basis of what makes one person different from another, encoding normal things like hair color. They are inherited and are present in all cells of the individual. This is in contrast to tumor specific mutations, which are errors, introduced in a single cell of an organism by, for example, irradiation. However, even though SNPs are inherited, normally occurring and in the whole body, they can still result in an increased risk for specific diseases like MM[65] and CLL. In the case of CLL, a SNP linked to the micro-RNAs mir-15a and mir-16-1 has been linked to familial CLL, and many other genes like IRF4 and LEF1, known to be aberrantly expressed in CLL, are shown to be associated with SNPs. This is not only important for inheritance, but also in trying to understand the disease, as supposedly unaffected tissue from the same patient, like nails, saliva or other available cells, is often used as control to look for tumor related mutations.

3.4.2 Recurrent mutations

One way of altering a gene's expression or effect is by mutating single base pairs from one base to another, altering the genetic code. Mutations can be coding or non-coding, respectively falling within or outside of the coding part of a gene. When a coding mutation results in an amino acid change of the coded protein, it is called non-synonymous. When the mutation has no effect, it is called silent. Silent mutations accumulate in normal cells by errors during cell division, or by damage inflicted by mutagens like UV radiation. However, cancer cells can have a particularly high load of mutations, as DNA repair and pathways that lead to programmed cell death are often compromised as part of a cell's transition to cancer.

Coding On average, a single CLL genome displays around 1000 mutations, and about 10-20 of these fall in coding regions and are non-synonymous[66,67]. For MM the number of mutations is reported to be around 7450, of which 35 non-synonymous[68]. On a patient population basis, most of these mutations are rare, with only a few mutations recurring in >15% of patients, and numbers vary between studies. Some of the most frequent mutations in CLL are SF3B1, ATM, TP53, NOTCH1 and MYD88[69] (**Figure 7**), while for MM some are KRAS, NRAS, FAM46C, DIS3 and TP53[70]. However, these mutations are not restricted to CLL or MM, and factors like age and treatment influence the observed frequencies. Even within a patient, variant allele frequency (VAF), or the frequency at which the mutation is found compared to the wild type allele, can be low, indicating the mutation is more likely to be related to progression than onset of the disease, if it is relevant at all.

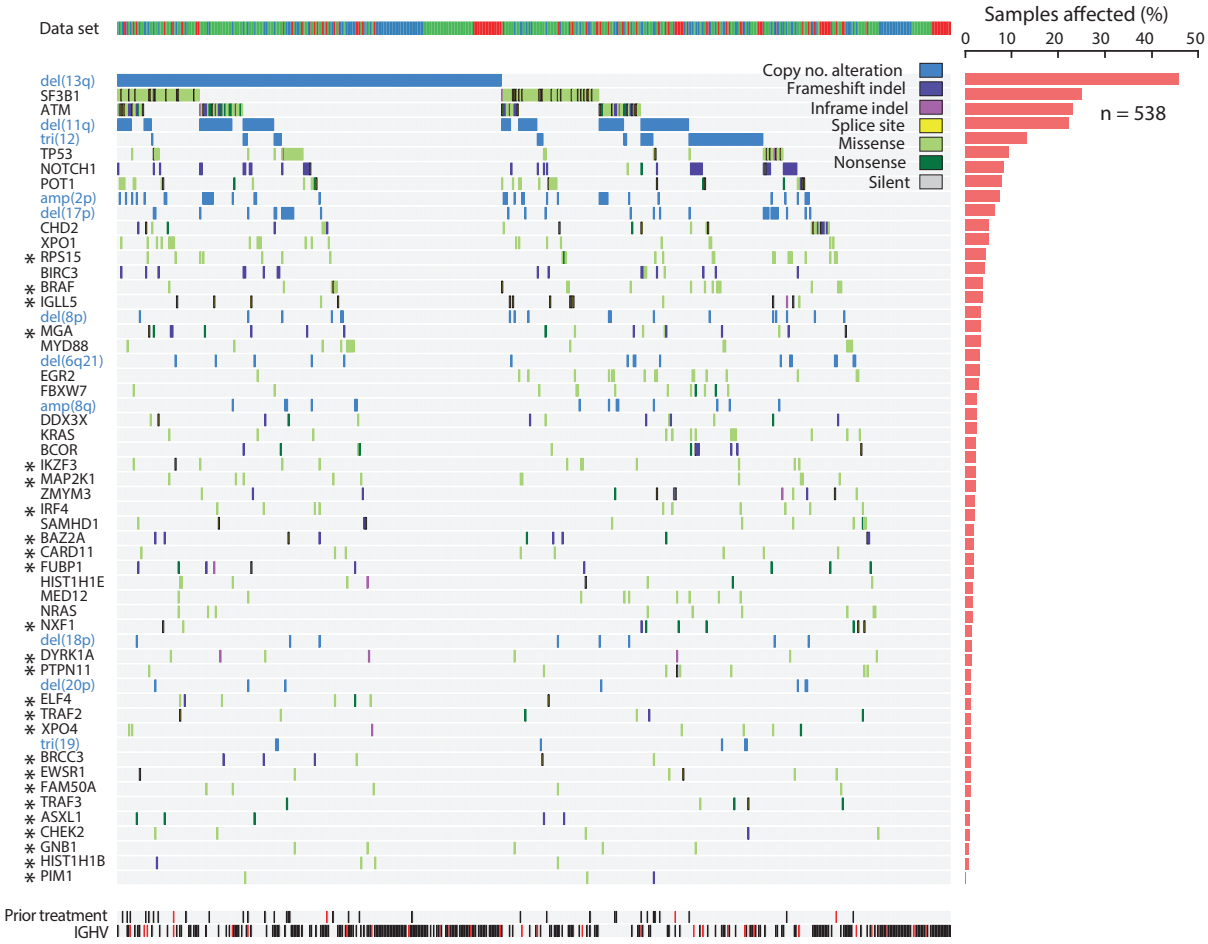


Figure 7: Frequencies of most common genetic events in CLL[69]. Summarizes data from the different datasets (top), different treatment history (black = treated, white = treatment naïve) and IGHV mutational status (white = U-CLL, black = M-CLL, red = unknown). Originally published by Landau et al., and reproduced with permission from Nature.

Non-coding Non-coding mutations provide an additional challenge, as putting them into their genetic context and finding possible effects is not straightforward. In CLL recurrent mutations have been identified in the 3' region of NOTCH1 and in an enhancer of PAX5[71]. A recent study in MM, using promoter capture Hi-C to link non-coding mutations to target regions, also describes PAX5 targeting, among others[72].

3.4.3 Chromosomal aberrations

Genetic changes can be a lot bigger than point mutations, as large parts of the genome can for example be deleted, duplicated or rearranged. These chromosomal aberrations, also called aCNV, are particularly frequent in cancer compared to normal cells and can affect multiple genes at once. Normally these changes would result in cell death, but the tumor suppressor genes needed for apoptotic programs are often among the ones affected in these cells.

CLL The four most common aCNVs in CLL are deletions on chr13 (del13q14), chr17 (del17p) or chr11 (del11q), and trisomy of chr12 (tri12) [67,73]. At diagnosis, del13q14 is found in about 50% of CLLs, making it the most frequent aCNV. Because of this, it is considered

important for CLL and considered a driver for pathogenesis. However, this identical deletion is also found in MM, other blood cancers and even solid tumors, making it unlikely that del13q14 causes CLL specific properties or is sufficient to cause CLL. The minimally deleted region includes two micro-RNAs, miR-15-a and miR-16-1, which are frequently down-regulated in CLL[74] and linked to TP53 and ZAP70[75]. The second most frequent aCNV at diagnosis is tri12 and is detected in about 15-18% of CLLs. Given that, unlike deletions, it is not bound to a very specific region in the genome, it is harder to study and the least understood of the recurrent aCNVs. It seems to be mutually exclusive with del13q14 and is present in roughly half of the NOTCH1 mutated cases. The third most frequent aCNV is del11q. It is detected in 10-15% of CLL cases upon diagnosis, results in the deletion of the ATM gene and nearly all cases belong to the U-CLL subgroup. The last frequent aCNV is del17p, present in 7-10% of CLL cases and also recurrently deleted in MM and other cancers. It always includes the deletion of TP53 on at least one allele, and the remaining copy is mutated in virtually all cases. About 80% of CLL patients have at least one of these four aCNVs [73], and although a range of less frequent aCNVs can occur, only about 20% of CLLs have 3 or more aCNVs. Thus, albeit diverse between patients, CLL is considered a rather stable disease.

MM As compared to CLL, MM can be very complex, even within a patient. In addition to del13q14 and del17p shared with CLL, MM also presents with common deletions on chromosomes 1, 6, 8, 11, 14 and 16 and gains on chromosomes 1, 12 and 17[76]. Also the absence or addition of whole sets of full chromosomes is common, leading to aberrant states of hyperhaploidy (24–34 chromosomes), hypodiploidy (35–45 chromosomes), hyperdiploidy (47–57 chromosomes) and hypotriploidy (58–68 chromosomes) [76–78]. Also translocations are common in MM, in which parts of the chromosomes are relocated. The two most frequent are t(11;14), and t(4;14), respectively present in 14% and 11% of patients[76]. Both of these events involve chromosome 14, as this is where IGH genes are located. Because of the importance of antibodies in the workings of B-cells, they are heavily expressed, and thus translocation to these regions can cause dramatic overexpression of the relocated genes, as they can highjack the strong enhancers present there. The fact that the regions need to be rearranged during normal B-cell development, to create a repertoire of antibodies, makes them a natural weak spot for errors, some of which may lead to MM[79].

3.5 EPIGENETIC CHANGES IN CLL & MM

With whole genome sequencing methods becoming more available, more and more studies look at the epigenetics behind CLL and MM. For this work, the main interest is the transcriptional regulation side of this research area, rather than, for example, DNA damage repair systems. As in the previous chapter, DNA methylation will be introduced for its central role in the concept of epigenetics, and histone modifications for their role in the papers included in this work.

3.5.1 DNA methylation

Altered DNA methylation is a general phenomenon in cancer[80]. It has long been recognized that cancers show general hypomethylation, a loss of overall methylation levels in the genome[81]. Given that methylated regions are silenced, a general loss of silencing results in aberrant activation of regions, making them more prone to instability, and opening the way to cancer[82]. To add insult to injury, specific sites, like promoters of tumor suppressor genes, get hypermethylated, silencing them and removing their ability to control the tumor[83].

CLL A lot of research effort has gone into describing and understanding DNA methylation in CLL. As with cancer in general, CLL presents with global hypomethylation[84]. More disordered methylation patterns are linked to worse outcome, and hypothesized to add another level of genetic instability, giving the CLL clone more options to gain fitness and expand[85]. In addition to general disruption, many regions consistently gain or lose methylation in CLL compared to healthy controls, and U-CLL and M-CLL have distinguishable methylation patterns in line with their assumed cells of origin[84,86]. In fact, DNA methylation studies suggest CLL is more of a disease continuum derived from any time point in the B-cell maturation process, rather than a binary disease derived from just naïve or memory B-cells[87].

MM The situation is very similar for MM, though some patients have been described to show global hypermethylation[88]. In line with the extensive list of genetic events, MM displays more heterogeneity in DNA methylation than other hematopoietic malignancies, including CLL[88]. Still, many recurrent changes are found, and a plethora of genes and pathways are shown to be affected by altered DNA methylation[89]. In the non-coding genome, binding sites of under-expressed B-cell TFs falling in enhancers were found to be hypermethylated, as if they had been decommissioned after the TF expression was down-regulated[88].

3.5.2 Histone modifications

Many aberrant histone modifications, be they global or local, have been described in cancer, leading to the suggestion of the term “histone onco-modifications”[90]. They include many different histone marks, and influence not just gene expression, but also genomic stability, DNA repair and cell cycle checkpoints.

CLL As with genetic changes, aberrant histone modifications are yet another way to disrupt genes in cancer. For example, the microRNAs involved in del13 can also be silenced by histone deacetylases (HDACs)[91]. These epigenetic modifiers are overexpressed in CLL, and removing them with an HDAC inhibitor (HDACi) is shown to kill CLL cells[92]. HDACs do not work on one site only, and thus it is not surprising that HDACi treatment influences the histone acetylation levels of many genes of interest in CLL[93]. On a more global level, a recent, large scale study published histone ChIP-seq for 107 CLL patients[94], significantly adding to the data available on this topic. In line with DNA methylation findings, U-CLL and M-CLL are distinguishable on H3K27ac level, with many regulatory regions being

selectively active in one or the other, showing varying states of overlap with different healthy B-cell populations.

MM Also in MM the genetics links back to the epigenetics. The gene NSD2 (also known as MMSET) for example, is the target of the t(4;14) translocation, resulting in its up-regulation in MM patients with this aberration. Since NSD2 is an epigenetic modulator, its up-regulation changes levels of H3K36 and H3K27 methylation, resulting in a more open state[95], which is in line with the global DNA hypomethylation. Like in CLL, HDACi are of interest in MM research, and in this case even approved for therapy[96], highlighting the importance of histone modifications. On a local scale, B-cell enhancers with aberrant DNA hypermethylation in MM were shown to have high levels of the expected histone modifications (H3K4me1, H3K27ac) and DNase sensitivity in a MM cell-line[88].

4 METHODOLOGICAL CONSIDERATIONS

This sections serves as a complement to the materials and methods in the included works. It presents a selection of topics, discusses the related methodological choices, and contrasts differences between the studies.

4.1 METHODS OVERVIEW

The following figures (**Figure 8 to Figure 11**) give a schematic overview of the projects, and my contribution to them (bright green = hands on, pastel green = varying degrees of providing code, coaching or discussion). For the sake of clarity, many side analyses, extra tests for reviewers and other work that did not make it into the final papers are not represented.

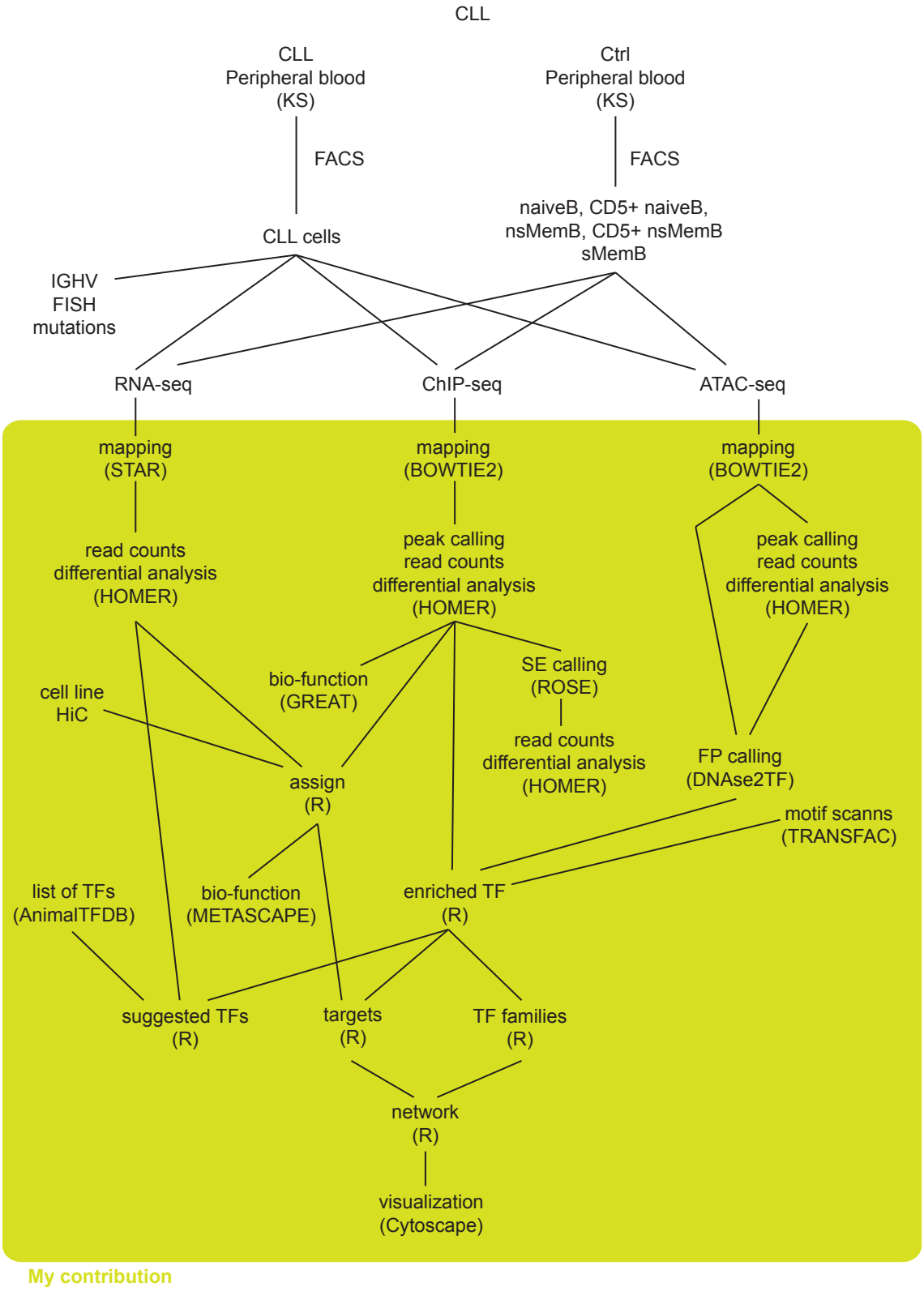


Figure 8: Overview of the CLL project.

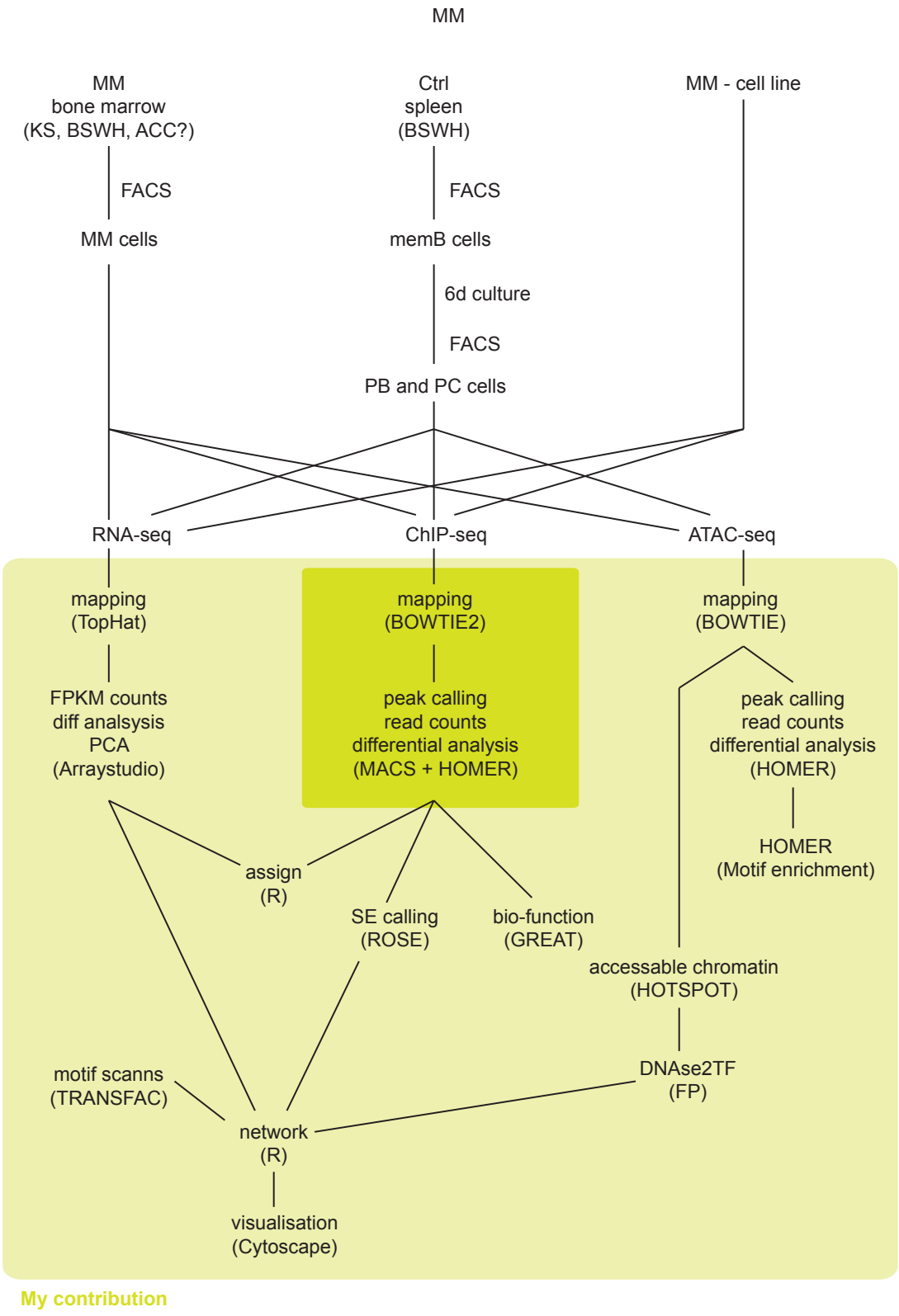


Figure 9: Overview of MM project.

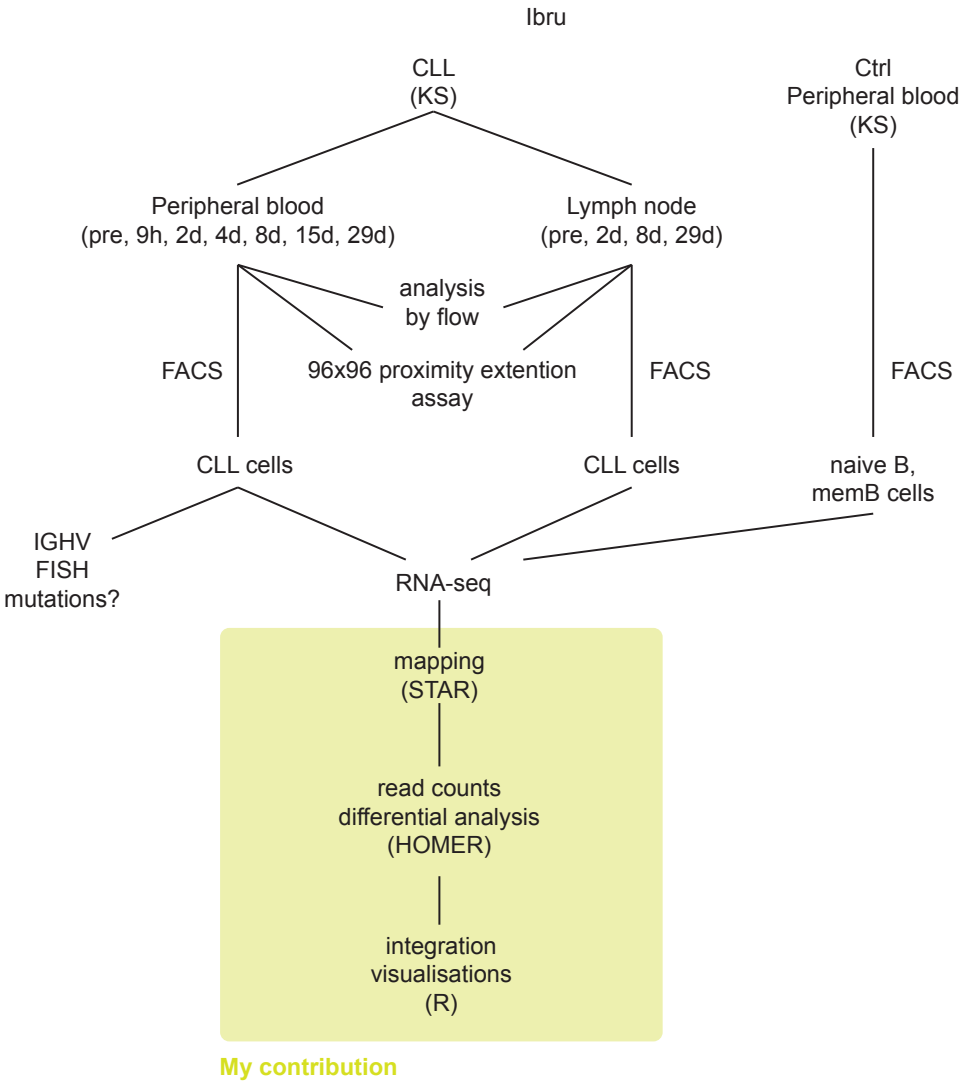


Figure 10: Overview of Ibru project.

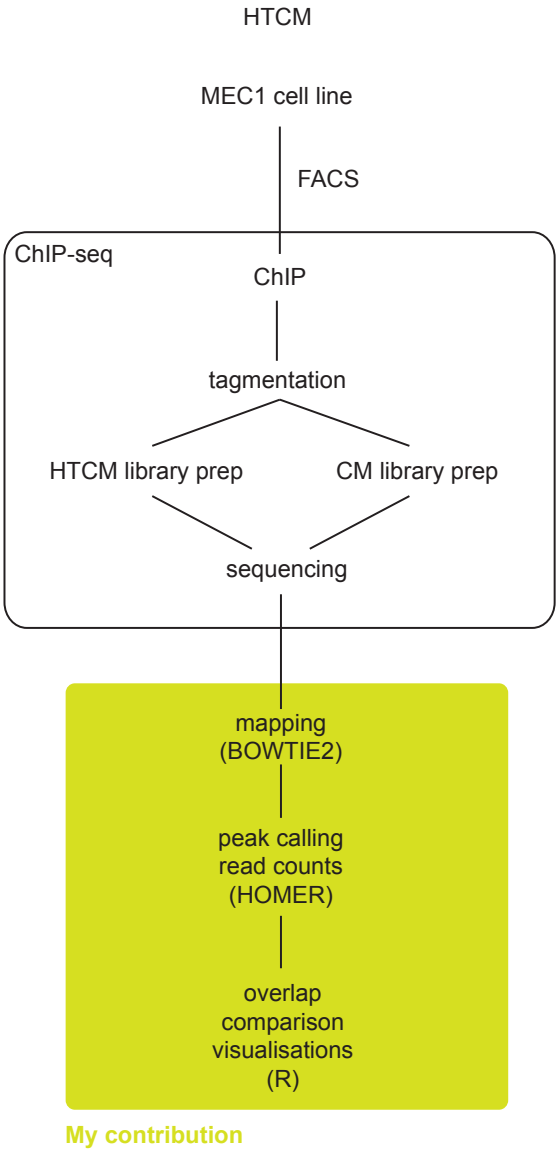


Figure 11: Overview of HTCM project.

4.2 PATIENT CHARACTERISTICS

As all studies are rather exploratory and small scale, the number of patients included in each study is small (**Table 3**). The number of normal controls is even more limited, but multiple normal B-cell populations were sorted from each donor, bringing the total number of control samples up, and covering both interpersonal and inter-population variation. In the HTCM study only the MEC1 cell line was used, so the only variation there was technical.

Table 3: Basic information of the samples included in the different projects.

	CLL		MM		Ibru		HTCM	
	case	ctrl	Case	ctrl	case	ctrl	case	ctrl
Total (n)	23	5	23	4	7	3		
Male (n)	12	4	14	1	6	?		
Female (n)	11	1	9	3	1	?	MEC1 cell line	
Treated?	mix	-	No	-	yes	-		
Age (y)	37-76	>60	49-91	NA	57-77	>60		

4.2.1 Sex and gender balance

The sex of the included subject was not part of the design. For patients in the CLL and MM studies, the balance is acceptable, especially given that both conditions are more common in men. For the Ibru study, men are over represented among patients. In case of healthy controls, which were de-identified before we received the samples, there was no possibility to know the sex of the subjects, until samples were sequenced and we could deduce sex from the read coverage on the Y chromosome (for which we found ChIP-seq data more effective than RNA-seq). Even then, we did not exclude samples based on this criteria, and simply used all available controls. The healthy controls from the Ibru project were also included in the CLL project. No enquiry was done with respect to the gender the subject identified with.

4.2.2 Geographical origin

The MM patients included in the study came from 3 different centers, two in the US and one in Sweden, while normal controls were from US only. No obvious bias was visible, so no specific analysis was performed to describe geographical differences. The fact that not only local patients were involved does make the dataset more reflective of real life. Patients in the CLL and Ibru studies, as well as healthy controls, were from Sweden. The HTCM study was performed on the MEC1 cell line, completely removing personal variation between samples. No enquiry was done with respect to the ethnical background of the subjects.

4.2.3 Prior treatment

The CLL study includes more personal variation when it comes to treatment, as some patients had been treated before, and others had not. Patients in the Ibru study had all been treated with other agents prior to Ibru treatment for this study, though the number of regimes varied. MM patients were all treatment naïve at the time of sampling, and so this study is most

representative of disease basics, without interference of prior treatment, However, it also means the findings might be less relevant to real-world patients that have undergone prior treatment regimes.

4.3 CHOICE OF METHODS AND MARKS

4.3.1 Cell sorting

One of the major strengths of the included work is that all samples were carefully FACS sorted, rather than, for example, solely bead enriched. This higher purity results in improved comparisons between populations later on, as it reduces false positive or false negative differences caused by contamination with other cells. Even in the HTCM study, where only the MEC1 cell line was used, samples were still sorted for live cells. It is important to keep in mind though that all handling of cells will stress them, and possibly cause regulatory changes.

4.3.2 Histone ChIP-seq marks

In the CLL and MM paper, histone ChIP-seq was to locate regulatory regions in all samples, assess their activity and identify differences between patients and healthy controls. For the CLL paper, we selected the histone marks H3K4me2, for its marking of enhancers and promoters[38], and H3K27ac to distinguish active from poised regulatory elements[38,97]. Using them together can give extra insight[39] and both marks are often present in research characterizing the histone landscape[13,38]. An even better distinction could have been made had we identified enhancers and promoters separately, by use of H3K4me1 and H3K4me3 instead of H3K4me2. However, given that the focus of the paper was on gene regulation, the most important information was to know the activity of the elements, and thus H3K27ac took center stage in the CLL paper, and was the only mark interrogated for the MM paper. In the HTCM study, H3K27ac was used to test the new method's performance on histone modifications, as the antibody has high ChIP efficiency and we have ample experience with it, reducing technical variation of sources other than the tested method itself.

4.3.3 High-throughput ChIPmentation

A clear limitation of regular ChIP-seq is the amount of input cells needed. In both the CLL and MM projects, we aimed to have RNA-seq, ChIP-seq and ATAC-seq from all included patients and controls. However, there is always a limited amount of sample available, and in case of the normal controls, sorting of multiple populations from the same sample further reduced available material. At the end of 2015, and alternative ChIP-seq method named ChIPmentation (CM), because of its use of tagmentation for ChIP-seq, was published[98]. We did not have this method set up in the lab during the data generation of the CLL and MM papers, but started using and developing it for other projects. By further optimizing the CM protocol, we developed an even faster and easier version: high-throughput ChIPmentation (HTCM). If the experiments for the CLL and MM studies would be performed now, we would use this protocol, which might have allowed us to create a more complete dataset from the samples that were available at the time.

4.3.4 ATAC-seq vs TF ChIP-seq

In order to build the regulatory networks in the CLL and MM study, we needed to identify the presence of TFs and connect them with their possible gene targets. We opted to use ATAC-seq instead of individual TF ChIP-seq. As ATAC-seq indicates open chromatin and unprotected DNA, it can be used to identify regions were TFs might be binding. While greatly simplifying lab work thanks to high throughput and universal data, the main drawback of ATAC-seq for this use is that it does not give a direct answer as to what is occupying the identified space, and additional info is needed to make a best guess of the specific TF(s) present (**Table 4**). To make a best guess of the occupying TF, we have used overlap with motif scans based on the TRANSFAC database. However, it still means the analysis depends on previously described and ChIPped TFs present in a database, making it vulnerable to technical errors and complications due to biological redundancy of the TF binding sites.

Table 4: Comparison between TF ChIP-seq and ATAC-seq for assessing TF binding.

TF ChIP-seq	ATAC-seq
X know only what you chip (low-throughput)	☑ know all occupied regions at once (high throughput)
X dependent on availability and quality of anti-body	☑ no anti-body optimizations needed
X TF presence is only info you get	☑ could be used to look at histone occupancy
☑ know that this exact TF is present there	X needs extra data to figure out what could be occupying the space

Keeping in mind the relative high-throughput of the developed HTCM method, large scale TF ChIP-seq is more realistic, though the availability of anti-bodies remains a limiting factor.

4.3.5 Whole ATAC-seq peaks vs footprinting

Small, protected regions can be found within ATAC-seq peaks. These are considered footprints (FPs), the actual locations were TFs are binding and thus very locally protect the DNA, which can be queried on a genome-wide scale[99].

The CLL paper originally relied on HOMER motif enrichment within the whole ATAC-seq peaks, but after the use of FP in the MM study, the approach was reconsidered. In the CLL dataset, the median ATAC-seq peak is 172 bp, while for FPs this is 18 bp, so reducing the searching space seems appropriate. In line with the MM paper, FPs were called from ATAC-seq data. However, sequencing depth of the ATAC-seq was considerably lower in the CLL paper, meaning we had to pool CLL and normal control samples to reach a depth suitable for footprinting. In the MM paper, samples were kept separately, but only ones with high sequencing depth were considered. Scanning the reference genome for TF motifs resulted in putative TF binding sites (TFBSs). For both the CLL and MM paper, a TF showing a TFBS with minimum 50% overlap with a FP was considered to bind that region.

Though it is desirable to be able to pinpoint more specifically which regions to consider, it adds yet another technical step in which assumptions and approximations can exclude important information.

4.3.6 DNA methylation

As most epigenetic data available for CLL and MM is DNA methylation, we decided to focus on the lesser-explored side, the histone modifications and through those the regulatory elements. This both to add other information, and to restrict the number of methods performed, both from a sample availability and a method availably perspective. However, given that the state and fate of a cell is influenced by the interplay of all of these marks and forms of regulation, the absence of paired DNA methylation data for our samples could definitely be seen as a gap. During the review process of the MM paper, changes in regulatory element activity were overlapped with published changes in DNA methylation[88], with a clear enrichment compared to random sampling. However, this is just a far approximation, and paired data, using the exact same patients and controls as in our studies, would have been a lot more informative.

Also other forms of epigenetic regulation, like micro-RNAs and long non-coding RNAs, were not studied.

4.4 BASIC DATA PROCESSING

Many small decisions are made when selecting samples or filtering data, often without proper acknowledgement or discussion in final publications. This section touches on some of these steps, giving more background as to why these decisions were made and their possible implications.

4.4.1 Filtering X and Y

Handling X and Y chromosomes in next generation sequencing poses difficulties. First, given the partial similarities between both chromosomes, mapping is complicated due to multiple alignments, even in female individuals, were it is clear that reads mapping to the Y chromosome are technical errors. Standard procedures of removing ambiguously aligned reads will thus be targeting the sex chromosomes to a higher extend than other chromosomes. Additionally, the studies included here have only small sample sizes, and the included patients and healthy controls are not balanced when it comes to sex. Though technical solutions help to resolve the first issue[100], the second would only be solved by the inclusion of extra, carefully selected subjects. Instead, we opted to remove both sex chromosomes to avoid false positive disease-related differences. However, both CLL and MM are more common in male patients, so sex differences do seem to be relevant to the diseases. This is important to keep in mind, and might limit the generalization capacities of the studies. On the other hand, it is in line with the aim to present the part of the disease that is consistent among all patients, no matter their genetic make-up, sex or age.

4.4.2 Normalization

The absence, presence and type of normalizations used can have extensive influences on the results and validity of data analysis and, in the worst case, lead to different interpretation of the findings. This section gives some more explanation and reasoning behind the different basic forms of normalization used in the included works. It is important to note that differential analysis, performed by EdgeR and mostly through HOMER, was always performed on raw counts, as advised by the software.

Correction for multiple testing RNA-seq, ChIP-seq and ATAC-seq datasets consist of many thousands of genes or peaks. For each of these, all samples (patients and controls) have counts, and thus for each gene or peak one can test if there is a statistical difference between cases and controls, resulting in a p-value. However, when putting all these thousands of tests together, p-values lose their meaning, and the total set no longer represents a 1% change (in case of $p<0.01$) that there actually is no difference between case and control, even if the test seems to find some. To remedy this one needs to correct for multiple testing. A standard correction is to use the FDR (False Discovery Rate), but in most cases we opted to go for the more conservative Bonferroni correction. In this case, each single p-value is multiplied by the total number of tests in the dataset, and then the cutoff of 0.01 is used on this corrected value.

Log transformation The standard approach for all data in the included works was to get raw counts for all samples (not normalized), and then normalize the counts tables afterwards in R. The first normalization is a log10 or log2 transformation. Log transformation of sequencing data generally makes the distributions more normal, which makes them easier to handle, especially with statistics. To avoid issues with counts of 0 reads turning to –Inf after transformation, +1 was added to all counts prior to log transformation. It is also common to do log2 transformations instead of log10, which gives the same change in the shape of the distribution. Log2 lends itself easily to use with log2 fold changes, but it can be harder to have a feeling for the actual number.

Quantile normalization The second step of standardization used was quantile normalization. Making the assumption that the majority of genes or regions do not change in expression, activity or accessibility, the overall distribution of the samples can be assumed similar. However, differences in for example sequencing depth can shift distributions, and thus similar number of reads can be incomparable between samples. Quantile normalization assumes that the distributions should be the same, and tries to pull the minimum, first quintile, medium, third quantile and maximum of different samples to the same level. However, it is important to keep in mind that, if the distributions are actually biologically different, this *introduces* bias by arbitrarily making them the same. However, we saw improvements in the biological interpretability of for example PCAs, making it likely that the normalization removed technical variation rather than biological (**Figure 12**). Of note, counts for SEs were not quantile normalized, since these are fewer and selected for regions that are exceptionally active or inactive in case and control, so no assumptions about overall stable distributions were made.

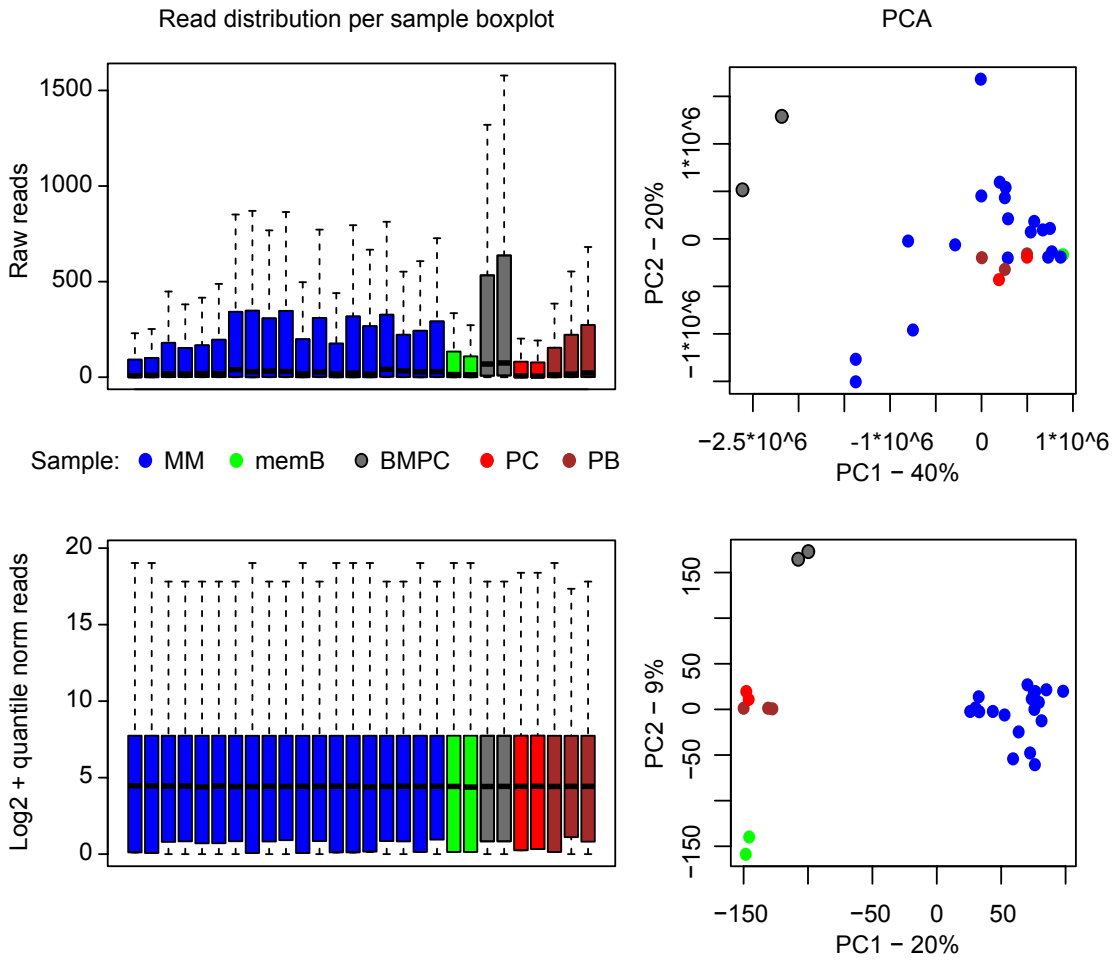


Figure 12: Raw and normalized reads for MM RNA-seq dataset, showing the improvement in read count distribution (left) and informative value of the data, as it reflects the biological populations (right).

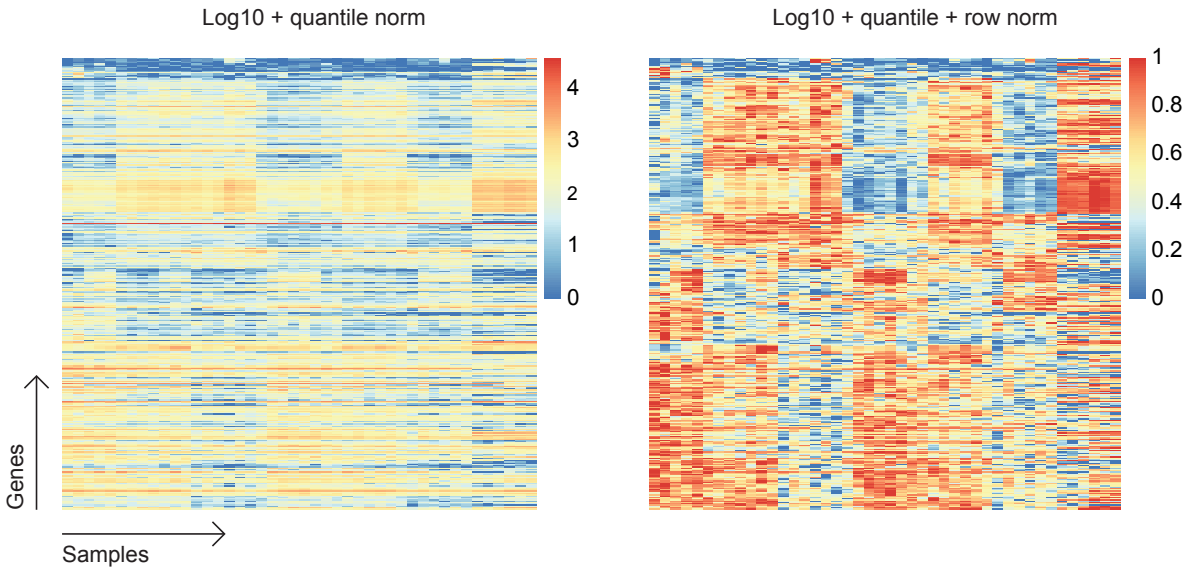


Figure 13: The effect of row normalization. Differentially expressed genes in Ibru dataset (**Figure 2A in Ibru**) illustrated as non-row normalized (left) and row normalized (right). Free clustering shows even more difference, but order of genes and samples used in the Ibru paper is preserved for ease of reference.

Row normalization In order to visualize genes or peaks with a wide range of read counts in a single plot, row normalization can help to show the differences one actually want to look at, which most often is the difference between samples, not between genes or peaks (**Figure 13**). It also influences the result in hierarchical clustering when plotting, and inherently removes any influence of gene or peak length. However, this means the info on the overall level of expression or activity is lost. In a row normalized heatmap, a gene changing from 5 reads to 10 reads will look the same as one changing from 1000 to 2000 reads.

4.4.3 Sample quality measures

For the CLL and Ibru project, samples were considered outliers if a technical explanation could be found. Firstly, fastq files were checked for general quality using FastQC, from which samples with big problems, like extreme duplication rates, were flagged, excluded from further analysis and redone if material was available. Secondly, samples not reaching a minimal number of processed reads (mapped and deduplicated) were re-sequenced or excluded. Cut-offs were 10M for RNA-seq and ATAC-seq and 25M for histone ChIP-seq. Thirdly, ChIP-seq and ATAC-seq samples with low peak enrichment were redone or excluded. Though this can be seen from visually inspecting tracks in a genome browser, we decided upon a more objective cut-off based on "Approximate IP efficiency" value of HOMER peak finding output, where a minimal value of 10 was required. Finally, read distributions were compared per data type (**Figure 14**), and samples with widely aberrant distributions were removed from analysis. All of this assumes that the included samples are not wildly different in, for example, DNA content of the cells. Given that all included cells are human B-cells, we made the assumption they would be very similar in general, and extreme aberrations as listed above would be purely technical. In the HTCM project this was not the case, as low read numbers or aberrant distributions give information on the library quality, which is exactly what we wanted to measure there.

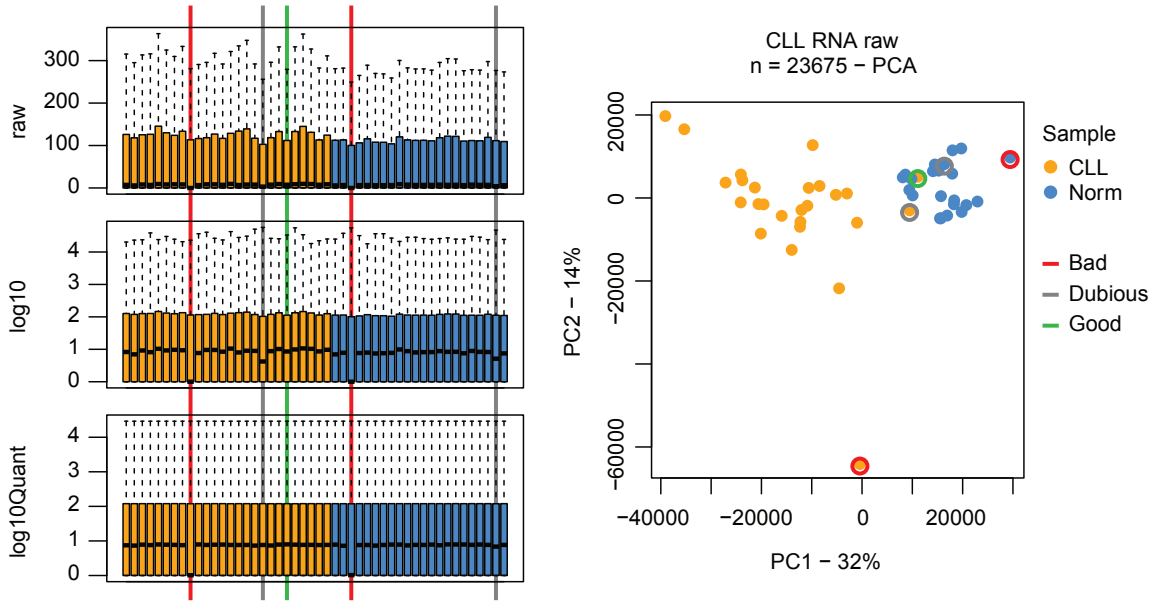


Figure 14: Sample quality of CLL RNA-seq. Read count distribution (left) and PCA (right).

4.4.4 Number of identified elements

Though it is natural to report the number of peaks, elements or genes one identifies or finds differential, it is a measurement that says just as much about the technological aspects of the data as about the biological ones. Even small differences in sample quality, lab efficiency, analysis settings etc. will give rise to different numbers. The influence of analysis decisions is exemplified in **Table 5**, which lists some H3K27ac ChIP-seq numbers for the CLL and MM studies, both having the same goal of looking at differences between healthy controls and the respective diseases.

Table 5: Difference in identified and differential peaks between studies and questions. All selections of differential peaks (diff) are based on Bonferroni corrected $p < 0.01$, except for the example of fold change (FC) selection, in which case peaks with a change of minimum 2 fold were selected.

H3K27ac peaks	total	diff	%diff
CLL vs ctrl	36594	3977	11%
MM vs memB	80407	1200	1%
(FC) MM vs memB	80407	20475	25%
MM vs (memB or PB)	80407	2006	2%
MM vs (memB and PB)	80407	201	0%

Peak total Already for the total number of identified peaks, in other words, all peaks in any of the samples, be they from patients or healthy controls, there is a big difference between the CLL and MM studies. This could be due to any kind of variation, be it inter-personal, biological or technical. It could be because MM is known as a more heterogeneous disease, because the normal controls cover a wider range of B-cell development, the difference in sample numbers or peak calling methods, etc. As a result, the absolute number of reported changes can not be compared, even if they had been called with the same software and calculated in the same way, which was not the case.

Use of controls As a start the CLL study considered all normal populations as one group of controls, asking for significant differences between all CLL and all controls. The MM study on the other hand kept the different populations of normal controls separate, so differences between MM and memB (memory B-cells, in the paper referred to as MB), PC (plasma cells) and PB (plasma blasts), were calculated separately. This might result in more stringency, but also a loss of statistical power, since it greatly reduces the number of controls for each question. So the 11% change in “CLL vs ctrl” compared to the 1% change in “MM vs MB” probably reflects the number of samples and the type of controls, rather than the biology.

Fold change Additionally, the MM study mostly selected changes based on fold change (FC) rather than statistical tests and p-values. This is a very big conceptual difference, and again makes the numbers incomparable, as the same “MM vs MB” question results in a 1% statistical change, but a 25% change based on FC. As we directly noticed during the study, selections based on FC are very venerable to systematic biases in, for example, sequencing depth. Prior to proper normalizations, this had lead the study to very different conclusions.

4.4.5 Implications for comparing methods

The HTCM paper assessed our improved version of the original ChIPmentation (CM) protocol[98]. This posed some specific issues, given that the difference in data quality between methods, and especially cell-numbers, was considerable. However, there was no point of removing low quality samples as described above, even if a technical explanation was readily found, as this is exactly what we wanted to describe.

Library complexity The difference between the methods lies in the library preparation, where HTCM results in reduced loss of material, and thus gives more complexity in a library from the same number of cells (**Figure 1F in HTCM**). However, sequencing one sample deeper than another also results in more duplicate reads. In order to be fair to the different methods, each sample from a given cell-number, no matter the method, was down-sampled to the number of reads found in the sample with lowest sequencing depth.

Normalization and filtering Most of the data handling was performed in line with the other project described in the sections above. However, HTCM and CM data was not quantile normalized, as we expect the read distributions to be widely different (**Figure S2D in HTCM**). Furthermore, no differential analysis was performed. Instead, we explored the number of detected peaks and their consistency between methods and cell-numbers.

Number of peaks Yet again, the number of identified peaks in the samples does not say everything. When looking at the samples in genome browser, we could clearly see that some samples were better than others, still they ended up with similar number of total reads. This is why we introduced the cutoff for peak quality, where the number of high quality peaks, as reported by HOMER, was plotted in addition to just to total number (**Figures S2B and S2C in HTCM**).

Peak overlap The fact that two methods find the same number of peaks does not mean that they find the same things. In line with the original CM paper, showing the overlap of peaks with regular ChIP-seq results, we described the overlap between our CM and HTCM results. Again, the focus was on high quality peaks, as they are the ones that we definitely want to retain between methods and cell-numbers (**Figures 1D and 1H in HTCM**).

4.5 DATA INTEGRATION

The CLL and MM papers make extensive use of data integration, given that they combine RNA-seq, ChIP-seq and ATAC-seq data. This sections highlights and contrasts some of the approaches.

4.5.1 Combining RNA-seq and histone ChIP-seq

The way regulatory elements were linked to their potential target genes differs between the MM and CLL papers.

CLL Promoters were defined as all H3K27ac peaks closer than +-2kb of the TSS of an expressed gene. For distal elements on the other hand, TADs (Topologically Associated

Domains) were used instead of CTCF-domains. Publically available HiC data from the GM12878 cell line was used to call TADs. Within these, correlation values between gene expression and H3K27ac signal were calculated, and used to inform the decision (**Figure 2 in CLL**). Elements were assigned to their closest expressed gene, but only if they were significantly positively correlated. If they were not, the element was assigned to the highest, significantly correlated gene within the TAD. If no correlations were positive and significant, elements were left unassigned. SEs were simply seen as a combination of individual elements, such that a SE was assigned to all genes its individual enhancer or promoter peaks were assigned to.

MM Regulatory elements closer than +-2 or 2.5kb were considered promoters. Distal regulatory elements were assigned based on a previous publication[101], assigning within so called CTCF domains. These domains extend +-200kb from the TSS of a gene, unless a CTCF site was crossed, in which case this became the domain border, as CTCF is assumed to work as an insulator. In contract to the original paper, the CTCF sites were not defined by ChIP-seq, but by overlapping CTCF motif scans with ATAC-seq peaks. Within these CTCF domains, elements that gained and lost activity were assigned to their closest gene. In the case of SEs, links were made by the ROSE software that called the SEs, and if a SE was linked to multiple genes, only the highest expressed was retained.

This includes a lot of differences. The MM approach has the advantage not to include cell line data and issues in TAD calling, but is dependent on databases, success of motif scanning and introduces an arbitrary distance cutoff of 200kb. Assigning elements to the closest gene possibly leaves fewer elements unassigned, but ignores the fact that they often skip over genes and target more distal ones, and while correlation does not mean causation, it might add info linked to function. For all of these reasons and more, a myriad of linking variations can be found in the literature.

4.5.2 Combining ATAC-seq and histone ChIP-seq

Direct overlap of regions is pretty straightforward compared to combining expression with regulatory regions. Yet simple choices are made along the way, which do have implications for the results.

CLL During the integration of ChIP-seq (H3K27ac and H3K4me2) and ATAC-seq peak catalogs, peaks were considered to overlap when the shared one base. It is to be questioned if this is biologically relevant, but given the relative uncertainty of the actual boundary of an element or binding location, we decided that it was enough. Since the reason for this overlap was to select regions open for binding in active regulatory elements, the main interest was the overlap between ATAC-seq and H3K27ac histone ChIP-seq. And surely, there are many of these elements. However, because we did full overlaps of all possible combinations of the three datasets, we could see that about 60% of ATAC-seq peaks do not fall within H3K27ac (**Figure S1B in CLL**). On top of that, the majority of those do not even fall within H3K4me2 peaks. None of these regions, though they can differentiate between CLL and healthy

controls (**Figure 15**), were used for the rest of the analysis, as they were not within the focus of active regulatory elements, as marked by H3K27ac. Additionally it should be mentioned that, apart from this exploration exercise, the ATAC-seq peaks themselves were never considered again, but rather the footprints falling within them.

MM In the MM paper, ATAC-seq peaks were only used by themselves. Integration with other datasets was only shown on footprint level, and focused on SEs rather than all H3K27Ac peaks.

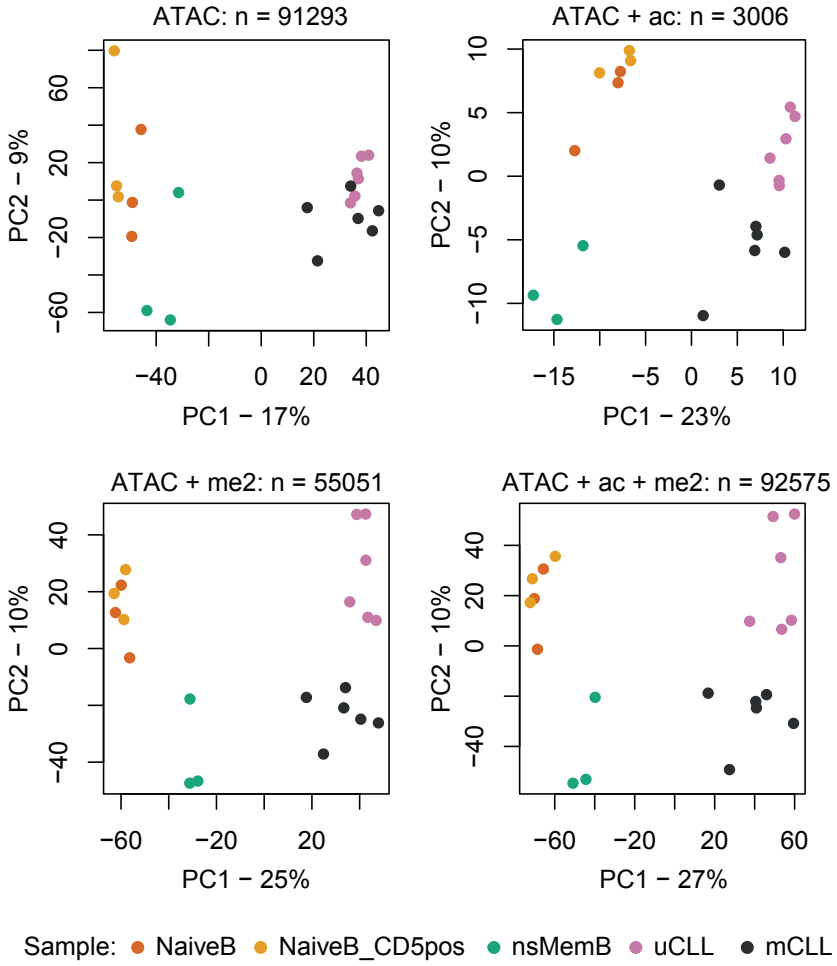


Figure 15: PCAs of CLL ATAC-seq based on overlap with H3K27ac (ac) and H3K4me2 ChIP-seq peaks (me2). Number of ATAC-seq peaks indicated above PCA.

4.5.3 Transcription factor enrichment

CLL Because we were initially looking for over-expressed TFs that clearly gained FPs, and under-expressed TFs that clearly lost FPs, enrichment was based on the difference between the number of gained and lost FPs that harbored a specific TF motif, and overlapped changing regulatory elements. Surprisingly, most TFs seem to have TFBSs falling in both gained and lost FPs, meaning that a low difference could point to either no changes, or very balanced changes. Additionally, absolute number of changes depends heavily on how easy it is to find the TFBS in the first place, as small motifs will have more hits over the genome, and thus will happen to fall in gained or lost FP more frequently than longer motifs do. For

this reason, the total number of TFBS for each TF was taken into account, no matter their overlap with FPs.

MM In contrast to the CLL paper, the MM paper only considered FPs within +/- 5kb from the TSS of the 542 TFs for which TFBS were available and which were linked to SEs. This reduced the complexity considerably, and as a result no further enrichment was needed.

4.5.4 Network approach

Both the CLL and MM paper include a regulatory network, which are the accumulation of the precious sections. Next to the differences described there, both papers consider different elements, target genes, and scope meaning that the resulting networks in the MM paper (**Figure 6 in MM**) and CLL paper (**Figure 7 in CLL**) are not directly comparable. In an attempt to contrast them anyway, a schematic overview can be found in **Table 6**. In summary, the CLL network is a global network focusing on differences between CLL and healthy controls, describing how TF (families) target genes through regulatory elements. The MM paper includes a core regulatory network, focusing only on MM, describing how TFs target each other through SEs.

Originally there was a second network in the MM paper, built from normal controls and constructed similarly to the MM diseased network, but it was not used in the publication. Instead, to contrast MM and healthy controls, simpler networks were built considering only regulation occurring around the TSS of the 542 included TFs, without additional information of regulatory elements.

Table 6: Comparison between CLL and MM network approaches.

	CLL	MM
Network type	Differential: network with changes between CLL and healthy subjects	Descriptive: network for MM only
Network scope	Global: regulatory elements to genes	Core regulatory circuit: SEs to TFs
Network construction	Single network build on integrated data from all patients and healthy controls	Integration of 8 networks build on data from 8 MM patients
Considered TFs	88 enriched TFs (based on differential FPs in differential regulatory elements)	TFs which were linked to SEs and for which TF motif scans were made
Considered regulatory elements	All differentially active regulatory elements	All distal SEs + SE regions +/- 5 kb around the TSS of the considered TFs
Considered target genes	All differentially expressed genes	The considered TFs
Considered FPs	All differential FPs (present in either CLL or healthy control) with TFBSs, falling in considered regulatory elements	All FPs with TFBSs, falling in considered regulatory elements
Solution to TF ambiguity	TF families, suggesting high and differentially expressed members	Random selection of highly expressed TF

5 RESULTS & DISCUSSION

Methodological topics, many of which can be considered to be findings in themselves, have already been touched upon in the previous chapter. This one summarizes the main points of the papers included in this work, and discusses their relevance.

5.1 CLL AND MM

As CLL and MM are very similar kinds of projects, they will be discussed together.

Consistent epigenetic patterns Both CLL and MM are genetically diverse, and MM even more so than CLL. Still, in both studies we found consistent changes in the use of regulatory elements, with thousands of elements showing consistently higher or lower activity in health than disease (**Figures 3A in CLL and 1B in MM**). On all levels available for the studies, be it expression (RNA-seq), establishment of regulatory elements (H3K4me2), activity of these elements (H3K27ac) or accessibility for TF binding (ATAC-seq), both CLL and MM datasets easily discriminated between healthy and diseased individuals (**Figures 1G in CLL and S2C, S3D and S5B in MM**). In the case of CLL, also the well-known clinically relevant subgroups of U-CLL and M-CLL were readily distinguished, and related to their supposed cells of origin (**Figure 5 in CLL**). We found ATAC-seq data to be particularly discriminatory and consistent with current classification (**Figure 1G in CLL**), so one might consider using this type of data for an epigenetic-based classification effort. Though the number of samples was limited, some discrimination between CLL with or without trisomy 12 could be found (**Figure 5A in CLL**). We also showed the use of SEs in different settings in both CLL (**Figure 4 in CLL**) and MM (**Figures 2, 3 and 6 in MM**), and the de-compaction of heterochromatin in MM (**Figure 4 in MM**).

Regulatory networks of disease In the CLL paper we show there is a vast amount of deregulation present in the wiring of CLL cells. The constructed networks suggest which TFs and TF families could be underlying the aberrant expression of hundreds of genes (**Figures 7B and 7C in CLL**), as highlighted in a selection of genes that are of great interest in CLL (**Figure 7D in CLL**). For example, we showed that LEF1, a gene known to be up-regulated in CLL, has many CLL specific enhancers (**Figure 1F in CLL**), and through these is differentially targeted by TFs (**Figures 7C and 7D in CLL**). As mentioned in the previous chapter, the nature of the MM (SE-TF) network is different, and it does not contrast results with healthy controls, thus showing the totality of the regulation in de MM B-cells, not merely the aberrant part. We show that, for example, IRF4 is central to the MM network and highly expressed (**Figure 6C in MM**) in addition to being linked to MM specific SEs (**Figure 2C in MM**). This shows its aberrant regulation, and that it subsequently may deregulate the cell further, accumulating in the network found in MM. Furthermore, comparison of the complexity of promoter-TF networks in individual MM patients and normal controls shows the loss of network complexity in MM (**Figure 5B in MM**).

Novelty For MM, the description of the thousands of consistently changed regulatory elements, based on RNA-seq, H3K27ac ChIP-seq and ATAC-seq, is novel information and

data, though maybe not surprising based on DNA methylation data and histone data from other malignancies. The de-compaction of heterochromatin is fully in line with the published observation of DNA hypomethylation in MM and cancer in general, which might make our observation more of a cancer conclusion than a MM specific concept. Still, it is a valuable additional resource now publically available for further exploration. For the CLL project, two very important papers have come out during the course of our research, which show a similar pattern, and include many more patients. The first was by Rendeiro et al,, publishing ATAC-seq for 55 patients[102]. The second was by Beekman et al., publishing not only ATAC-seq, but also H3K27ac ChIP-seq on 107 CLL patients for which DNA methylation data was already available, and comparing it to healthy B-cells[94]. Their bigger dataset allowed them to not only see a tri12 specific pattern, but also changes present in case of specific recurrent mutations. From a reproducibility perspective this is great news, as it is in agreement with what we have found, and shows the general conclusions hold true on a bigger scale, and are not just a fluke of our dataset. The question remains how big the overlap between their and our stably changed elements is, which also holds true for the vast amount of DNA methylation data published on both disorders.

Discussion After endless hunting for a genetic change that initiates all of the CLL or MM cases, and thus is present in all disease cases, it is very compelling to say that, among the consistent epigenetic patterns, we, meaning all researchers, have finally found the underlying mechanism for the diseases. However, there is nothing in this data that assures, or maybe even insinuates, that there is a causative relationship between the observed pattern and the initiation of the disease itself. What we are looking at is a relatively stable state, integrating everything from predisposition to initiation and progression, and in the case of CLL even different prior treatment regimes. So it might well be that, just to say it with a proverb, “all roads lead to Rome”. Still, it is plausible that random errors not only accumulate in the genome, but also the epigenome, and that multiple specific combinations of them ultimately result in one or the other disease. In light of this it would be interesting to see not only what is common between different patients or cohorts of the same disease, but also to overlap the CLL and MM data, and see how they relate to one another.

Relevance The value of these works lies in the fact that they form great resources for further exploration. For example, the thousands of differential regulatory regions can be used to assess the possible importance of non-coding mutations, and, when adding the network information to this, even what the mutation’s effect could be. Similarly, it might be used to see if a gene recurrently targeted by deletion or mutation, but is genetically normal in a given patient, might be targeted by epigenetic deregulation instead. Combining all that data, all the ways in which genes can be recurrently targeted, could then lead to finding actual hits present in all patients with a disease, just on different levels that so far could not be tested, or were not integrated. Instead of genes being reported as x% recurrently mutated, we could add how often they are deleted, translocated, targeted by DNA methylation, histone modifications, micro-RNAs etc. This might well lead us to the original step on the figurative road to Rome. The networks could also be used by researchers working on a specific gene or TF, and could

suggest effects of removing any of the TFs and breaking the network in light of therapy. Integrated networks, build in similar ways for different diseases, could give insight in the workings of currently used drugs, and their transferability to other diseases.

5.2 IBRU

Early response We measured inflammatory cytokines in plasma, presence of cell-surface markers by flow cytometry and gene expression in lymph nodes (LN) and peripheral blood (PB) at multiple time points, ranging from before to 29 days (d29) after treatment with Ibrutinib. Already 9h after treatment, levels of multiple inflammatory markers are significantly altered in plasma, most of them showing reductions, with changes lasting at least until d29 post-treatment (**Figure 1 in Ibru**). Additionally, most of the expression changes in both PB and LM observed between pre-treatment and d29 samples are already visible from the earliest time point, in this case two days (d2) after treatment (**Figure 2 in Ibru**). Increase of total circulation lymphocyte counts is visible at 9h, but only significant from the d2 time point (**Figure 3A in Ibru**).

Changes in LN last longer than in PB We show that the effect of Ibrutinib on expressional changes in LN stays stable or progresses further between d2 and d29 after treatment for almost all of the affected genes (**Figure 2 in Ibru**). However, though PB shows the same effects on d2, many of these changes revert back to pre-treatment levels by d29.

Novelty Previous studies have shown reduction of cytokine levels in blood plasma upon longer periods of treatment with Ibrutinib, for example after 8 weeks[103]. We confirm these findings, and add information on the speed of the changes, as assessment of levels just hours after treatment with Ibrutinib is new. Similarly, the expression of selected genes has been assessed on time points similar to the ones in this work[52], but RNA-seq, giving a much broader and unbiased view on the expression had not been done.

Discussion Though we as humans would always like to have quick working solutions to our medical problems, if it is anything more than a headache we usually do not get what we want. However, on a cellular and molecular scale, things do happen fast. Upon treatment with Ibrutinib, patients can report an almost instant improvement, which, based on what is known about the phenomenon of redistribution lymphocytosis combined with our results, can definitely be more than just the thought of getting treatment. The total lymphocyte and CLL counts, combined with the literature, suggest that the RNA-seq pattern we observe, the stability of change in LN and the initial similar change of PB, which then reverts back to the original PB state, is exactly this process of redistribution lymphocytosis. From the total blood counts we can see that this happens within hours, and the dramatic changes within the first two days observed in RNA-seq suggest the same. It does not seem unreasonable to think that this redistribution is the reason for the improved well being reported by patients. In hindsight it would haven been better to have even earlier time points, as clearly a lot of the changes in plasma levels already happened before 9h, and in expression levels before d2. However, all changes observed in this study, both between compartments and between time points, are a

lot smaller than the changes observed between healthy and diseased samples in the CLL study.

Relevance It is hard to argue that understanding more about the workings and effects of a drug currently used in treatment of patients could ever *not* be relevant. Be it for improved treatment regimes, handling of side-effects, its potential for combination therapy, its tendency to cause resistance or the development of new compounds, unraveling the relative mystery of why a lot of drugs work at all could be of great benefit. Also for basic research, studies of treatment are basically perturbation studies in the exact relevant complex system of a human body. This can help the understanding of complex concepts like relevance of the microenvironment, the mobility of cells and how they differ based on the compartment they are in, and so on.

5.3 HTCM

Faster We show that the original ChIPmentation (CM) protocol can be sped up by replacing the standard reverse crosslinking, which is often performed overnight, and DNA purification, adding another 1-2h, with a 5min step at 95°C in the PCR reaction that normally follows these procedures (**Figures 1I and S1 in HTCM**). We also show that, by reducing incubation times, the ChIP procedure itself can be shortened, making the whole HTCM procedure less than 24h (**Figure 1I in HTCM**), sequencing time included, without obvious loss of quality (**Figures 1B and 1G in HTCM**).

Lower cell-numbers By eliminating the pre-amplification DNA purification, the HTCM protocol loses less material than the CM one. This results in more complex and better quality libraries from the same cell numbers (**Figures 1B, 1F and S2 in HTCM**). Consequently, HTCM allows for good quality data from lower starting cell-numbers than CM, with very good performance down to 2500 cells for the tested antibodies. At that level, 80% of the high-quality peaks present in the reference sample (150k) are still detected, while CM needs between 10k and 50k cells, so about 10x more, to maintain the same retention of high quality peaks from its reference sample (**Figure 1H in HTCM**).

Novelty The suggested HTCM protocol dramatically speeds up the originally published CM method[98], in addition to realizing better quality libraries from lower cell-number due to its reduction in loss of material. After we had set up the method, we discovered that the suggested PCR based reverse crosslinking approach was used before in another variation on ChIP-seq, called LobChIP[104]. However, though it reduces time and losses at the reverse crosslinking stage, it does not include the advantages of tagmentation, and thus the overall process is not as fast and simple as HTCM.

Discussion On multiple occasions, the CLL and MM paper highlight differences between primary samples and cell lines (GM12878 and MM.1S respectively), yet the HTCM paper was done on the MEC1 cell line. However, for this methodological proof of concept it was more important to have an unlimited amount of standardized cells, so the results could easily be reproduced by others, or compared to existing data. In hindsight it might have been

interesting to test the method on K562 cells instead of MEC1, as that was the cell line used in the original ChIPmentation paper. As mentioned in the chapter on methodological considerations, the HTCM method was not available at the time when the data for the CLL and MM papers was generated. Had it been, less cells of the available samples would have been needed for the generation of H3K27ac and H3K4me data, and we might have been able to generate complete data (RNA-, ChIP- and ATAC-seq) for more patients. Obviously, the use of this method does not solve the unspecific binding of low quality antibodies, so it does not quarantine that 2500 cells will be enough for any mark, on any cell type.

Relevance With the use of next generation sequencing methods in clinical applications lurking around the corner, fast, high throughput and low input methods are very attractive. The speed means that results can be available the day after a clinical visit, and the scalability ensures that many patients, or many different ChIPed marks per patient, could be tested in parallel. Of course, this is not only useful in clinical practices, but also in basic or translational research. As illustrated by the CLL and MM studies, being able to get high quality data from fewer cells means more can be done with the same sample. Also for rare cell populations this is good news, as a couple of thousand cells can be enough for a decent HTCM sample. Additionally, the fact that it is easy to multiplex means screening for the binding of different TFs is a lot more feasible. In general, improving methods opens the door to new applications and research possibilities.

6 CONCLUSIONS

The highlights of the different papers included in this work, as discussed in the previous chapter, are summarized in **Figure 16**.

In the end, this journey is an exploration of what computational biology can mean for biological systems. The result is only ever maximally as good as its input. In this case it means that there is a focus on patients from highly developed countries with similar lifestyles, and that gender differences remain unexplored, possibly reducing the power for generalization. Yet, hopefully bioinformatic efforts like this can reduce the number of tests that need to be performed in the lab by providing some acceptable best guesses, especially in light of avoiding wasteful use of samples freely given by patients, or unnecessary breeding and sacrificing of animals.

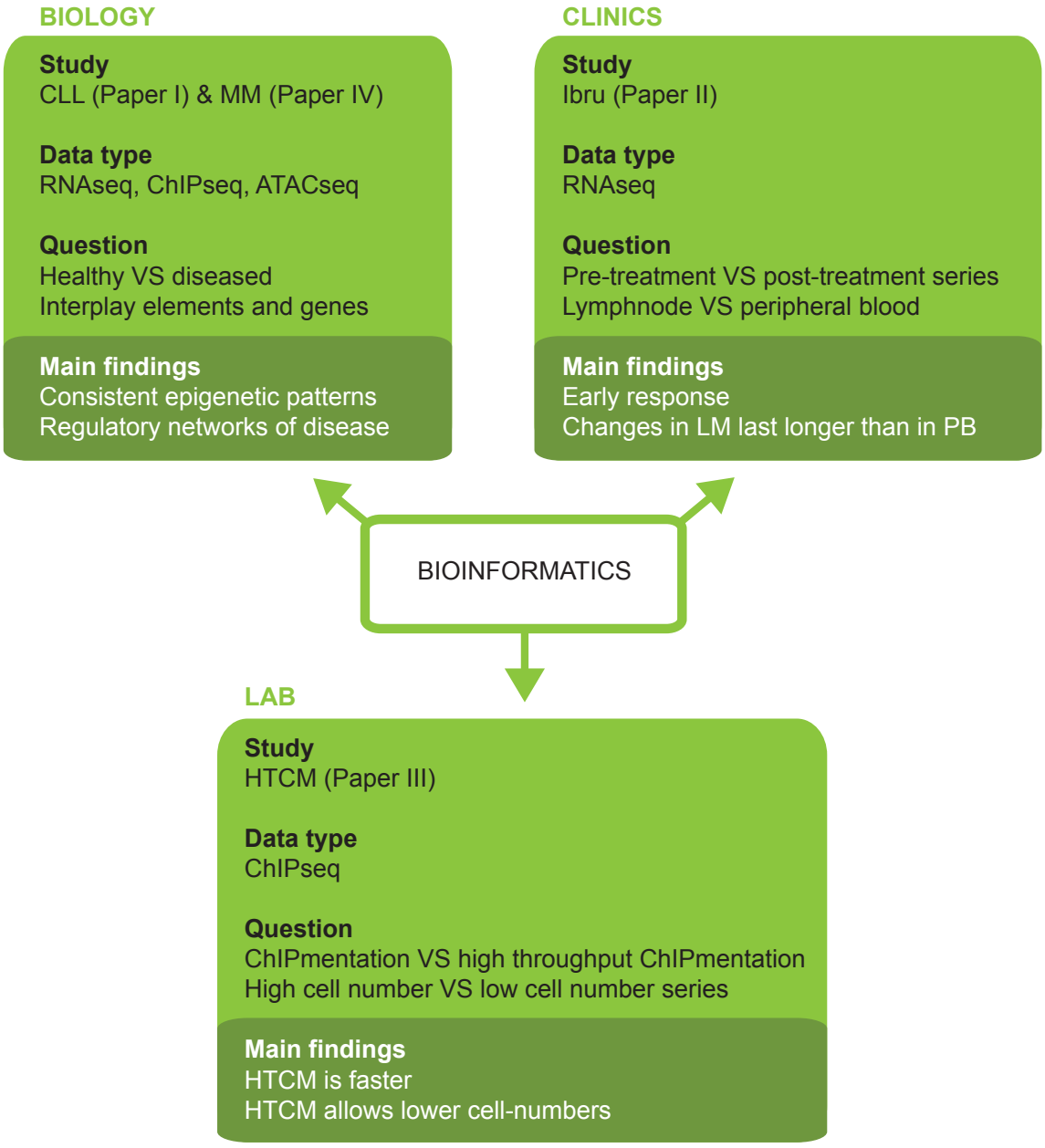
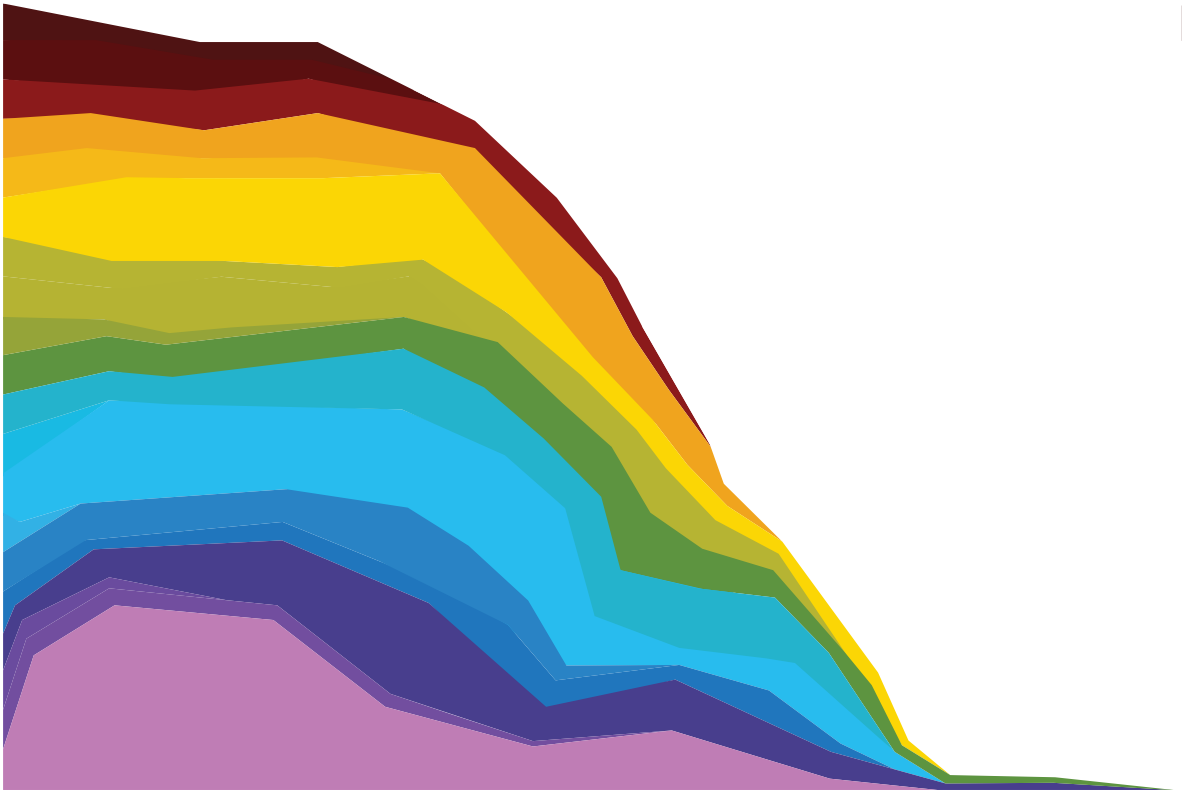


Figure 16: Summary of the different papers included in this work.



Cover: Artistic interpretation of the loss of diversity in the B-cell repertoire upon expansion of a malignant clone.
By Ayla De Paepe

7 POPULAR SCIENCE SUMMARY

Every person has a very specific genetic code, present in all cells of your body and holding all the information that makes you… You. But still, you have a lot of different cell types. How does a stomach cell know that it should produce digestive acids, while a skin cell knows it shouldn't? Why don't the cells in your ears produce saliva? This is where gene regulation and epigenetics come in! It tells a cell which parts of its genetic code should be used at which times, and which ones it can totally ignore (like how the part on "how to produce saliva" is ignored by the cells in your ears).

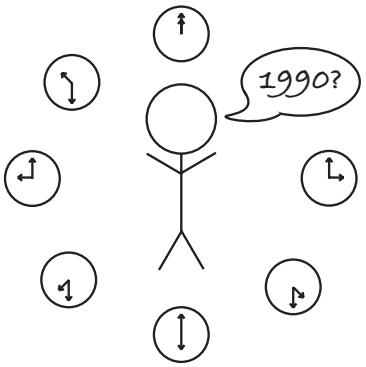
One of the less visible but important systems in your body is your immune system. The task of the immune system is to defend you against infections. It is made up of a bunch of different cell types circulating in your blood and keeping and eye on the rest of your body. Bye the way, did you know that all blood cells are made in the center of your bones? Pretty awesome if you ask me! Well, one specific set of immune system cells are the B-cells. They are the cells that get activated when you get a vaccination, and are exactly the ones that are studied in all of the papers in this thesis. When they get triggered, either by an actual infection or a vaccination, they help to clear the intruders from your body and build up a memory for this specific disease. In case you ever encounter it again, your body will be ready to get rid of the intruders so efficiently that you might not even notice you were infected at all. But, as with practically all cells in the body, sometimes things go wrong, and the cells can keep on growing and growing and become a cancer. When this happens to B-cells, there are multiple types of cancer you can get, and we have looked at two: chronic lymphocytic leukemia (CLL in short) and multiple myeloma (MM in short). The cancer cells grow so fast that they take over all the B-cell variation you are supposed to have (as shown in the colorful picture on the cover!).

So, what have we found by doing all this research? In two of the studies included in this thesis we show that there are a whole bunch of regions in the genetic code of sick B-cells that get activated or deactivated in the diseases we studied (paper 1 and paper 4). Knowing these regions could help not just our understanding of regulation, but hopefully also give an idea on how to make the sick cells normal again, or how to make them disappear. We also looked at the changes in active genes and molecules in the blood after treatment of patients with one of the drugs currently used in the hospital. With many drugs, we know they work, and might have an idea on what the mechanism is, but don't really know what physically happens to the body and the cells when people get treated. We saw there are some very early effects, were cells change which genes are active and move around in the body just hours or days after the patients got the drug (paper 2). And the last project included here is one where we suggest a faster and better method to look at some of the epigenetic changes, opening more doors for researchers to understand how cells work (paper 3). That's it!

8 ACKNOWLEDGEMENTS

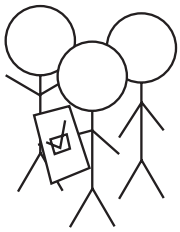
There are many people I would like to thank. If you are reading this, you are surely one of them, even if you cannot find your name, so: Thank you. And for all to read, I thank…

Robert



for accepting a 90´s kid into your lab, dealing with my stubbornness and all the time you put into my supervision

Pelin, Julian and Lennart

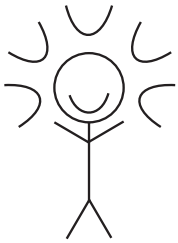


for being there when needed most, be it for technical, mental or signature support :)

Catarina

for being the best mentor ever! The fact that I made it is proof

Charlotte

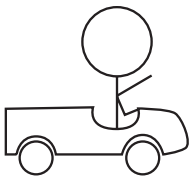


for being the light of the lab and all the important conversations

Lucia

for being the new queen of our data analysis

Yin Lin's lab



for the collaboration and your drive to get things out

Thibault

for being the joker

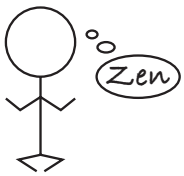
Minna

for your great social-ness

Shabnam

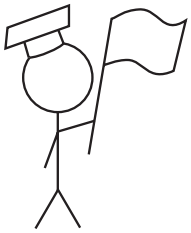
for your kind heart

Aleksandra



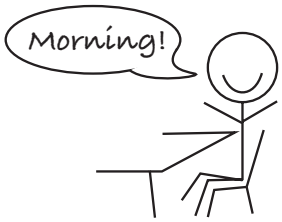
for the endless patience with the CLL paper

DSA and Karolina



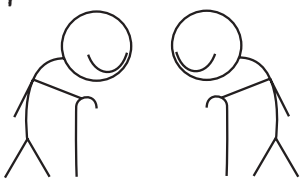
for reviving the student representative in me

Admin and everyone I've ever shared office with



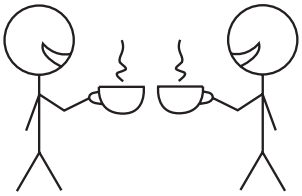
for the daily interactions

The PhD and post-doc oldies



for being around since I started!

Juan, Monika, Huthayfa, Aditya, Hani, Caroline

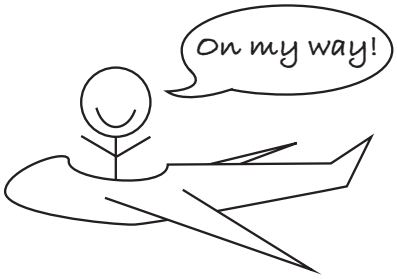


for the coffees,
the laughs and the tears

KICS, JPIAMR and bloggers

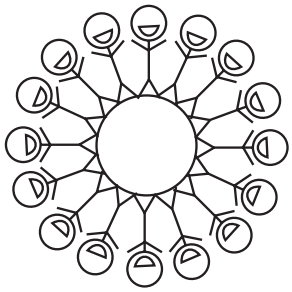
for the good times and
the amazing chance
to do an internship

Valérie



for all the visits,
the persistent friendship
and the pensées positives!

The Niyla crowd

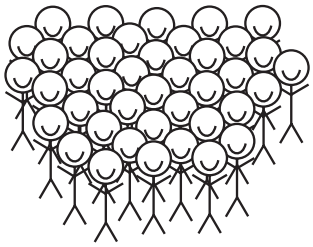


for being such
an awesome bunch!

My family

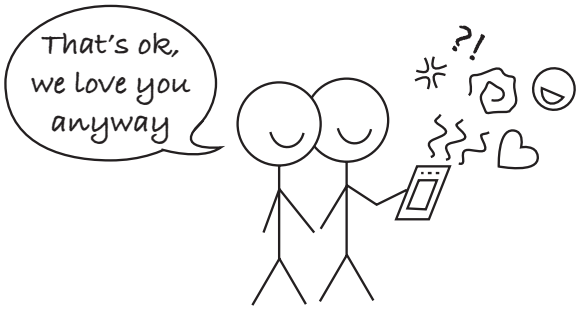
for sticking with me
from a distance

Nigel's family



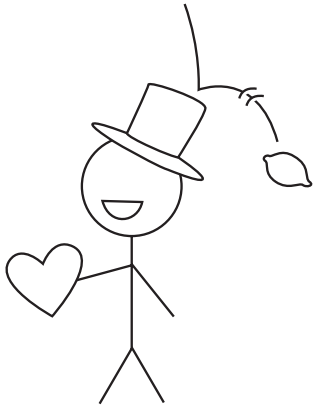
for welcoming me into your lives

Mams and Paps



for letting me move all over the place,
for carrying the worries of the past years,
for loving me

Nigel
my partner, my love



for being my white hat
in times when all that I can see
is just another lemon tree

(Yes, that is a reference to the song "Lemon tree" by Fool's Garden.
I felt I had to make that clear, since this is a scientific work and referencing is important)

REFERENCES

1. Rad A. Hematopoiesis (human) diagram en - File:Hematopoiesis (human) diagram.svg - Wikimedia Commons. https://commons.wikimedia.org/wiki/File:Hematopoiesis_(human)_diagram.svg#/media/File:Hematopoiesis_(human)_diagram_en.svg. Accessed November 26, 2018.

2. Tangye SG, Good KL. Human IgM+CD27+ B Cells: Memory B Cells or "Memory" B Cells? *J Immunol*. 2007;179:13-19. doi:10.4049/jimmunol.179.1.13

3. Maston G a, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet*. 2006;7:29-59. doi:10.1146/annurev.genom.7.080505.115623

4. Smale ST, Kadonaga JT. The RNA polymerase II core promoter. *Annu Rev Biochem*. 2003;72(1):449-479. doi:10.1146/annurev.biochem.72.121801.161520

5. Spitz F, Furlong EE. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet*. 2012;13(9):613-626.

6. Pott S, Lieb JD. What are super-enhancers? *Nat Genet*. 2014;47(1):8-12. doi:10.1038/ng.3167

7. Whyte WA, Orlando DA, Hnisz D, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*. 2013;153(2):307-319. doi:10.1016/j.cell.2013.03.035

8. Hnisz D, Abraham BJ, Lee TI, et al. Super-enhancers in the control of cell identity and disease. *Cell*. 2013;155(4):934-947. doi:10.1016/j.cell.2013.09.053

9. Mansour MR, Abraham BJ, Anders L, et al. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science (80-)*. 2014;346(6215):1373-1377. doi:10.1126/science.1259037

10. Chapuy B, McKeown MR, Lin CY, et al. Discovery and Characterization of Super-Enhancer-Associated Dependencies in Diffuse Large B Cell Lymphoma. *Cancer Cell*. 2013;24(6):777-790. doi:10.1016/j.ccr.2013.11.003

11. Lové J, Hoke HA, Lin CY, et al. Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers. *Cell*. 2013;153:320-334. doi:10.1016/j.cell.2013.03.036

12. Pabo CO, Sauer RT. Transcription Factors: Structural Families and Principles of DNA Recognition. *Annu Rev Biochem*. 1992;61(1):1053-1095. doi:10.1146/annurev.biochem.61.1.1053

13. Lara-Astiaso D, Weiner A, Lorenzo-Vivas E, et al. Chromatin state dynamics during blood formation. *Science (80-)*. 2014;345(6199):943-949. doi:10.1126/science.1256271

14. Subrahmanyam R, Du H, Ivanova I, et al. Localized epigenetic changes induced by DH recombination restricts recombinase to DJH junctions. *Nat Immunol*. 2012;13(12):1205-1212. doi:10.1038/ni.2447

15. Su I, Basavaraj A, Krutchinsky AN, et al. Ezh2 controls B cell development through

histone H3 methylation and Igh rearrangement. *Nat Immunol*. 2003;4(2):124-131. doi:10.1038/ni876

16. Velichutina I, Shaknovich R, Geng H, et al. EZH2-mediated epigenetic silencing in germinal center B cells contributes to proliferation and lymphomagenesis. *Blood*. 2010;116(24):5247-5255. doi:10.1182/blood-2010-04-280149

17. Good-Jacobson KL, Chen Y, Voss AK, Smyth GK, Thomas T, Tarlinton D. Regulation of germinal center responses and B-cell memory by the chromatin modifier MOZ. *Proc Natl Acad Sci*. 2014;111(26):9585-9590. doi:10.1073/pnas.1402485111

18. Li G, Zan H, Xu Z, Casali P. Epigenetics of the antibody response. *Trends Immunol*. 2013;34(9):460-470. doi:10.1016/j.it.2013.03.006

19. King RC, Stansfield WD, Mulligan PK. *A Dictionary of Genetics*. Oxford University Press; 2006. doi:10.1093/acref/9780199766444.001.0001

20. Riggs AD, Martienssen RA, Russo VE. *Epigenetic Mechanisms of Gene Regulation*. Cold Spring Harbor Laboratory Press; 1996.

21. Chandler VL. Paramutation's Properties and Puzzles. *Science (80-)*. 2010;330(6004):628-629. doi:10.1126/science.1191044

22. Bird A. Perceptions of epigenetics. *Nature*. 2007;447(7143):396-398. doi:10.1038/nature05913

23. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet*. 2003;33(3s):245-254. doi:10.1038/ng1089

24. Stricker SH, Köferle A, Beck S. From profiles to function in epigenomics. *Nat Rev Genet*. 2016;18(1):51-66. doi:10.1038/nrg.2016.138

25. Bernstein BE, Meissner A, Lander ES. The Mammalian Epigenome. *Cell*. 2007;128(4):669-681. doi:10.1016/j.cell.2007.01.033

26. Ecker JR, Bickmore W a., Barroso I, Pritchard JK, Gilad Y, Segal E. Genomics: ENCODE explained. *Nature*. 2012;489(7414):52-55. doi:10.1038/489052a

27. Ehrlich M, Gama-Sosa MA, Huang LH, et al. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. *Nucleic Acids Res*. 1982;10(8):2709-2721. doi:10.1093/nar/10.8.2709

28. Scarano E, Iaccarino M, Grippo P, Parisi E. The heterogeneity of thymine methyl group origin in DNA pyrimidine isostichs of developing sea urchin embryos. *Proc Natl Acad Sci U S A*. 1967;57(5):1394-1400. doi:10.1073/pnas.57.5.1394

29. Deaton AM, Bird A. CpG islands and the regulation of transcription. *Genes Dev*. 2011;25(10):1010-1022. doi:10.1101/gad.2037511

30. Bird AP. CpG islands as gene markers in the vertebrate nucleus. *Trends Genet*. 1987;3(C):342-347. doi:10.1016/0168-9525(87)90294-0

31. Antequera F, Bird a. Number of CpG islands and genes in human and mouse. *Proc Natl Acad Sci U S A*. 1993;90(24):11995-11999. doi:10.1073/pnas.90.24.11995

32. Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the

human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A*. 2006;103(5):1412-1417. doi:10.1073/pnas.0510310103

33. Ball MP, Li JB, Gao Y, et al. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat Biotechnol*. 2009;27(4):361-368. doi:10.1038/nbt.1533

34. Yisraeli J, Frank D, Razin A, Cedar H. Effect of in vitro DNA methylation on beta-globin gene expression. *Proc Natl Acad Sci*. 1988;85(13):4638-4642. http://www.pnas.org/content/85/13/4638.abstract.

35. McKeon C, Ohkubo H, Pastan I, de Crombrugghe B. Unusual methylation pattern of the alpha 2 (l) collagen gene. *Cell*. 1982;29(1):203-210. doi:0092-8674(82)90104-0 [pii]

36. Jenuwein T, Allis CD. Translating the histone code. *Science (80-)*. 2001;293(5532):1074-1080. doi:10.1126/science.1063127

37. Heintzman ND, Stuart RK, Hon G, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*. 2007;39(3):311-318. doi:10.1038/ng1966

38. Encode Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74. doi:10.1038/nature11247

39. Wang Y, Li X, Hu H. H3K4me2 reliably defines transcription factor binding regions in different cells. *Genomics*. 2014;103(2-3):222-228. doi:10.1016/j.ygeno.2014.02.002

40. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. *CA Cancer J Clin*. 2018;68(1):7-30. doi:10.3322/caac.21442

41. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394-424. doi:10.3322/caac.21492

42. Ferlay J, Ervik M, Lam F, et al. Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. https://gco.iarc.fr/today. Published 2018. Accessed November 19, 2018.

43. Paintmaps: Create Custom Map Charts with Free, Online Map Maker. Color Maps Online. https://paintmaps.com/. Accessed November 28, 2018.

44. Kipps TJ, Stevenson FK, Wu CJ, et al. Chronic lymphocytic leukaemia. *Nat Rev Dis Prim*. 2017;3:16096. doi:10.1038/nrdp.2016.96

45. American Cancer Society: Key Statistics for Multiple Myeloma. https://www.cancer.org/cancer/multiple-myeloma/about/key-statistics.html. Published 2018. Accessed November 23, 2018.

46. de Claro RA, McGinn KM, Verdun N, et al. FDA Approval: Ibrutinib for Patients with Previously Treated Mantle Cell Lymphoma and Previously Treated Chronic Lymphocytic Leukemia. *Clin Cancer Res*. 2015;21(16):3586-3590. doi:10.1158/1078-0432.CCR-14-2225

47. Deeks ED. Ibrutinib: A Review in Chronic Lymphocytic Leukaemia. *Drugs*.

2017;77(2):225-236. doi:10.1007/s40265-017-0695-3

48. Chanan-Khan A, Cramer P, Demirkan F, et al. Ibrutinib combined with bendamustine and rituximab compared with placebo, bendamustine, and rituximab for previously treated chronic lymphocytic leukaemia or small lymphocytic lymphoma (HELIOS): a randomised, double-blind, phase 3 study. *Lancet Oncol*. 2016;17(2):200-211. doi:10.1016/S1470-2045(15)00465-9

49. Byrd JC, Brown JR, O'Brien S, et al. Ibrutinib versus Ofatumumab in Previously Treated Chronic Lymphoid Leukemia. *N Engl J Med*. 2014;371(3):213-223. doi:10.1056/NEJMoa1400376

50. Burger JA, Tedeschi A, Barr PM, et al. Ibrutinib as Initial Therapy for Patients with Chronic Lymphocytic Leukemia. *N Engl J Med*. 2015;373(25):2425-2437. doi:10.1056/NEJMoa1509388

51. Herman SEM, Gordon AL, Hertlein E, et al. Bruton tyrosine kinase represents a promising therapeutic target for treatment of chronic lymphocytic leukemia and is effectively targeted by PCI-32765. *Blood*. 2011;117(23):6287-6296. doi:10.1182/blood-2011-01-328484

52. Herman SEM, Mustafa RZ, Gyamfi JA, et al. Ibrutinib inhibits BCR and NF-κB signaling and reduces tumor proliferation in tissue-resident cells of patients with CLL. *Blood*. 2014;123(21):3286-3295. doi:10.1182/blood-2014-02-548610

53. Herman SEM, Niemann CU, Farooqui M, et al. Ibrutinib-induced lymphocytosis in patients with chronic lymphocytic leukemia: correlative analyses from a phase II study. *Leukemia*. 2014;28(11):2188-2196. doi:10.1038/leu.2014.122

54. Kumar SK, Rajkumar V, Kyle RA, et al. Multiple myeloma. *Nat Rev Dis Prim*. 2017;3:17046. doi:10.1038/nrdp.2017.46

55. Nabhan C, Aschebrook-Kilfoy B, Chiu BC-H, et al. The impact of race, ethnicity, age and sex on clinical outcome in chronic lymphocytic leukemia: a comprehensive Surveillance, Epidemiology, and End Results analysis in the modern era. *Leuk Lymphoma*. 2014;55(12):2778-2784. doi:10.3109/10428194.2014.898758

56. Cerhan JR, Slager SL. Familial predisposition and genetic risk factors for lymphoma. *Blood*. 2015;126(20):2265-2273. doi:10.1182/blood-2015-04-537498

57. Schinasi LH, De Roos AJ, Ray RM, et al. Insecticide exposure and farm history in relation to risk of lymphomas and leukemias in the Women's Health Initiative observational study cohort. *Ann Epidemiol*. 2015;25(11):803-810. doi:10.1016/j.annepidem.2015.08.002

58. Baumann Kreuziger LM, Tarchand G, Morrison VA. The impact of Agent Orange exposure on presentation and prognosis of patients with chronic lymphocytic leukemia. *Leuk Lymphoma*. 2014;55(1):63-66. doi:10.3109/10428194.2013.794267

59. Waxman AJ, Mink PJ, Devesa SS, et al. Racial disparities in incidence and outcome in multiple myeloma : a population-based study. *Blood*. 2015;116(25):5501-5507. doi:10.1182/blood-2010-07-298760.The

60. Halvarsson B, Wihlborg A, Ali M, et al. Direct evidence for a polygenic etiology in familial multiple myeloma. *Blood Adv*. 2017;1(10):619-623. doi:10.1182/bloodadvances.2016003111

61. Khuder SA, Mutgi AB. Meta-analyses of multiple myeloma and farming. *Am J Ind Med*. 1997;32(5):510-516. doi:10.1002/(SICI)1097-0274(199711)32:5<510::AID-AJIM11>3.0.CO;2-5

62. American Cancer Society: Signs and Symptoms of Chronic Lymphocytic Leukemia. https://www.cancer.org/cancer/chronic-lymphocytic-leukemia/detection-diagnosis-staging/signs-symptoms.html. Published 2018. Accessed November 23, 2018.

63. Rajkumar SV. Multiple myeloma: 2018 update on diagnosis, risk-stratification, and management. *Am J Hematol*. 2018;93(8):1091-1110. doi:10.1002/ajh.25117

64. American Cancers Society: Signs and Symptoms of Multiple Myeloma. https://www.cancer.org/cancer/multiple-myeloma/detection-diagnosis-staging/signs-symptoms.html. Published 2018. Accessed November 23, 2018.

65. Mitchell JS, Li N, Weinhold N, et al. Genome-wide association study identifies multiple susceptibility loci for multiple myeloma. *Nat Commun*. 2016;7:12050. doi:10.1038/ncomms12050

66. Puente XS, Pinyol M, Quesada V, et al. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature*. 2011;475(7354):101-105. doi:10.1038/nature10113

67. Malek S. *Advances in Chronic Lymphocytic Leukemia*. Springer; 2013. doi:10.1007/978-1-4614-8051-8

68. Chapman MA, Lawrence MS, Keats JJ, et al. Initial genome sequencing and analysis of multiple myeloma. *Nature*. 2011;471(7339):467-472. doi:10.1038/nature09837

69. Landau DA, Tausch E, Taylor-Weiner AN, et al. Mutations driving CLL and their evolution in progression and relapse. *Nature*. 2015;526(7574):525-530. doi:10.1038/nature15395

70. Walker BA, Boyle EM, Wardell CP, et al. Mutational Spectrum, Copy Number Changes, and Outcome: Results of a Sequencing Study of Patients With Newly Diagnosed Myeloma. *J Clin Oncol*. 2015;33(33):3911-3920. doi:10.1200/JCO.2014.59.1503

71. Puente XS, Beà S, Valdés-Mas R, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature*. 2015;526(7574):519-524. doi:10.1038/nature14666

72. Hoang PH, Dobbins SE, Cornish AJ, et al. Whole-genome sequencing of multiple myeloma reveals oncogenic pathways are targeted somatically through multiple mechanisms. *Leukemia*. 2018;32(11):2459-2470. doi:10.1038/s41375-018-0103-3

73. Artmut H, Öhner D, Tephan S, et al. Genomic aberrations and survival in chronic lymphocytic leukemia. *N Engl J Med*. 2000:1910-1916.

74. Calin GA, Dumitru CD, Shimizu M, et al. Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc Natl Acad Sci*. 2002;99(24):15524-15529. doi:10.1073/pnas.242606799

75. Fabbri M, Bottoni A, Shimizu M, et al. Association of a MicroRNA/TP53 Feedback Circuitry With Pathogenesis and Outcome of B-Cell Chronic Lymphocytic Leukemia. *Am Med Assoc*. 2011;305(1):59-67.

76. Morgan GJ, Walker BA, Davies FE. The genetic architecture of multiple myeloma. *Nat Rev Cancer*. 2012;12(5):335-348. doi:10.1038/nrc3257

77. Sawyer JR, Tian E, Shaughnessy Jr JD, et al. Hyperhaploidy is a novel high-risk cytogenetic subgroup in multiple myeloma. *Leukemia*. 2017;31(3):637-644. doi:10.1038/leu.2016.253

78. Walker BA, Boyle EM, Wardell CP, et al. Mutational Spectrum, Copy Number Changes, and Outcome: Results of a Sequencing Study of Patients With Newly Diagnosed Myeloma. *J Clin Oncol*. 2015;33(33):3911-3920. doi:10.1200/JCO.2014.59.1503

79. Gonza D, Burg M Van Der, Fenton JA, et al. Review in translational hematology Immunoglobulin gene rearrangements and the pathogenesis of multiple myeloma. *Blood*. 2016;110(9):3112-3122. doi:10.1182/blood-2007-02-069625.

80. Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. *Carcinogenesis*. 2009;31(1):27-36. doi:10.1093/carcin/bgp220

81. Feinberg AP, Gehrke CW, Kuo KC, Ehrlich M. Reduced genomic 5-methylcytosine content in human colonic neoplasia. *Cancer Res*. 1988;48(5):1159-1161. http://www.ncbi.nlm.nih.gov/pubmed/3342396. Accessed November 16, 2018.

82. Eden A, Gaudet F, Waghmare A, Jaenisch R. Chromosomal Instability and Tumors Promoted by DNA Hypomethylation. *Science (80-)*. 2003;300(5618):455. doi:10.1126/science.1083557 ARTICLE

83. Herman JG, Baylin SB. Gene Silencing in Cancer in Association with Promoter Hypermethylation. *N Engl J Med*. 2003;349(21):2042-2054. doi:10.1056/NEJMra023075

84. Kulis M, Heath S, Bibikova M, et al. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat Genet*. 2012;44(11):1236-1242. doi:10.1038/ng.2443

85. Landau DA, Clement K, Ziller MJ, et al. Locally Disordered Methylation Forms the Basis of Intratumor Methylome Variation in Chronic Lymphocytic Leukemia. *Cancer Cell*. 2014;26(3):813-825. doi:10.1016/j.ccell.2014.10.012

86. Queirós a C, Villamor N, Clot G, et al. A B-cell epigenetic signature defines three biological subgroups of chronic lymphocytic leukemia with clinical impact. *Leukemia*. 2015;(29):598-605. doi:10.1038/leu.2014.252

87. Oakes CC, Seifert M, Assenov Y, et al. DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. *Nat Genet*. 2016;48(3):253-264. doi:10.1038/ng.3488

88. Agirre X, Castellano G, Pascual M, et al. Whole-epigenome analysis in multiple myeloma reveals DNA hypermethylation of B cell-specific enhancers. *Genome Res*. 2015;25(4):478-487. doi:10.1101/gr.180240.114

89. Amodio N, D 'aquila P, Passarino G, Tassone P, Bellizzi D. Epigenetic modifications in multiple myeloma: recent advances on the role of DNA and histone methylation. *Expert Opin Ther Targets*. 2017;21(1):91-101. doi:10.1080/14728222.2016.1266339

90. Füllgrabe J, Kavanagh E, Joseph B. Histone onco-modifications. *Oncogene*.

2011;30(31):3391-3403. doi:10.1038/onc.2011.121

91. Sampath D, Liu C, Vasan K, et al. Histone deacetylases mediate the silencing of miR-15a, miR-16, and miR-29b in chronic lymphocytic leukemia. *Blood*. 2012;119(5):1162-1172. doi:10.1182/blood-2011-05-351510

92. El-Khoury V, Moussay E, Janji B, et al. The Histone Deacetylase Inhibitor MGCD0103 Induces Apoptosis in B-Cell Chronic Lymphocytic Leukemia Cells through a Mitochondria-Mediated Caspase Activation Cascade. *Mol Cancer Ther*. 2010;9(5):1349-1360. doi:10.1158/1535-7163.MCT-09-1000

93. Jordaan G. Identification of Histone Epigenetic Modifications with Chromatin Immunoprecipitation PCR Array in Chronic Lymphocytic Leukemia Specimens. *J Cancer Sci Ther*. 2014;06(09):325-332. doi:10.4172/1948-5956.1000290

94. Beekman R, Chapaprieta V, Russiñol N, et al. The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. *Nat Med*. 2018;24(6):868-880. doi:10.1038/s41591-018-0028-4

95. Martinez-Garcia E, Popovic R, Min DJ, et al. The MMSET histone methyl transferase switches global histone methylation and alters gene expression in t(4;14) multiple myeloma cells. *Blood*. 2011;117(1):211-220. doi:10.1182/blood-2010-07-298349

96. Yee AJ, Raje NS. Panobinostat and Multiple Myeloma in 2018. *Oncologist*. 2018;23(5):516-517. doi:10.1634/theoncologist.2017-0644

97. Creyghton MP, Cheng AW, Welstead GG, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci*. 2010;107(50):21931-21936. doi:10.1073/pnas.1016071107

98. Schmidl C, Rendeiro AF, Sheffield NC, Bock C. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nat Methods*. 2015;12(10):963-965. doi:10.1038/nmeth.3542

99. Neph S, Vierstra J, Stergachis AB, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*. 2012;489(7414):83-90. doi:10.1038/nature11212

100. Wise AL, Gyi L, Manolio TA. EXclusion: Toward integrating the X chromosome in genome-wide association analyses. *Am J Hum Genet*. 2013;92(5):643-647. doi:10.1016/j.ajhg.2013.03.017

101. Akhtar-Zaidi B, Cowper-Sal-lari R, Corradin O, et al. Epigenomic enhancer profiling defines a signature of colon cancer. *Science*. 2012;336(6082):736-739. doi:10.1126/science.1217277

102. Rendeiro AF, Schmidl C, Strefford JC, et al. Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-specific epigenome signatures and transcription regulatory networks. *Nat Commun*. 2016;7(May):11938. doi:10.1038/ncomms11938

103. Niemann CU, Herman SEM, Maric I, et al. Disruption of in vivo chronic lymphocytic leukemia tumor-microenvironment interactions by ibrutinib - Findings from an investigator-initiated phase II study. *Clin Cancer Res*. 2016;22(7):1572-1582. doi:10.1158/1078-0432.CCR-15-1965

104. Wallerman O, Nord H, Bysani M, Borghini L, Wadelius C. lobChIP: from cells to sequencing ready ChIP libraries in a single day. *Epigenetics Chromatin*. 2015;8(1):25. doi:10.1186/s13072-015-0017-5